Statistical Methods for Analyzing the Built Environment

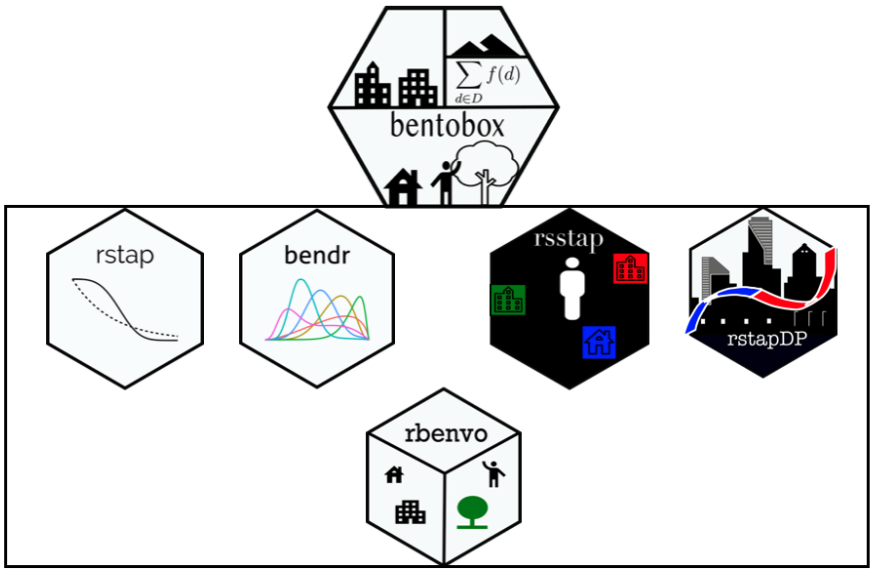
by

Adam Peterson

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2021

Doctoral Committee:

Professor Trivellore E. Ragunathan, Co-Chair
Professor Brisa N. Sánchez, Co-Chair
Professor Jian Kang
Professor Marie S. O'Neill


bentobox
rstap
bendr
rsstap
rstapDP
rbenvo

Adam Peterson
atpvyc@umich.edu
ORCID: 0000-0001-7071-7873
© Adam Peterson 2021
All Rights Reserved

Dedicated to my family, friends and teachers.

ACKNOWLEDGEMENTS

First and foremost I would like to thank, acknowledge, and generally express gratitude for my phenomenal advisor Professor Brisa Sánchez who has had to put up with, what I am sure has been, one of the most peculiar PhD training experiences she has ever had. Whether planning for how to work with me during a global pandemic or patiently listening to me as I regale her with my previous week's adventures, she has provided a consistently positive perspective and force in my life for the past four years. Secondly, I would be remiss if I did not did not similarly thank my committee, Professors Trivellore Ragunathan, Marie O'Neil and Jian Kang for their insightful and constructive comments on my dissertation work.

Next, there are a number of professors at Michigan for whom I remain thankful for playing a part in my education – statistical and otherwise. Professors Ananda Sen and Hyun Min Kang were both instrumental in supporting my interest in Statistics early on and in the PhD program specifically. I do not think I would have arrived at this moment without their encouragement and support. On that same note, Professors Rod Little and separately Jian Kang, Zhenke Wu and Hui Jiang each exposed me to a wide array of statistics literature in their class and journal clubs' respectively, that greatly broadened my view of the field. While not directly colleagues in statistical matters, I had the privilege of working alongside Professors Bhramar Mukherjee, Tom Braun, Mike Elliot, and Mike Boehnke on efforts to maintain and improve the student experience in the Department of Biostatistics. Their dedication to the department and their students is inspiring and they deserve all the praise they receive.

I am grateful to have had several friends who have supported me throughout my graduate school education with their wonderful company and friendship. These include, in no particular order and amongst others, Adam and Madeline Younkin, Sami and Ben Daniels, Chris Gore, Paul H., Jesse Day, Kelly Speth, Juan Marquez, Michelle Ngo, Desmond Kearsley, Andrew Whiteman, Leo Tse, Emily Hector, Emily Plieusch and Kevin Putschko. Special thanks to Andrew, Emily Hector and Holly Hartmann for sitting with me in hospital when I lost consciousness during a JSM conference and special thanks to Elizabeth Chase, Fatema Khorasani, Lauren Beesley, Tian Gu and Pedro Orozco del Pino for working alongside me on student experience projects.

Finally and most importantly, I must acknowledge my family, who have been a constant source of support and joy in my life. My love, thanks and gratitude to my parents: Tom Peterson and Deborah Pascal, as well as to my sister's family: Kelly and Stephen Kipp and my two wonderful nieces Eliza and Juliette. Whether near or far during graduate school, I always cherished my place as son, brother, and uncle.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF ABBREVIATIONS

- BEF: Built Environment Feature
- HFS: Healthy Food Store
- FFR: Fast Food Restaurant
- STAP: Spatial Temporal Aggregated Predictor
- DP: Dirichlet Process
- STAP-DP: Spatial Temporal Aggregated Predictor - Dirichlet Process
- MESA: Multi-Ethnic Study of Atherosclerosis
- NDP: Nested Dirichlet Process
- BKMR: Bayesian Kernel Machine Regression
- CGLM: Consensus Generalized Linear Model

ABSTRACT

The built environment refers to the human made space in which humans live, work and recreate on a day-to-day basis. As such, it constrains and enables individual choices and has consequently received increased attention for its potential influence on human health. For example, the availability of junk food outlets near children's schools may influence child obesity and the availability of supermarkets near residential addresses of study participants may influence longitudinal change in body mass index. However, efforts to estimate these influences have met methodological obstacles including the need to address residential self-selection bias and challenges related to defining measurement of environmental attributes. In response to these challenges, this work develops three statistical methods that seek to characterize the health effect of the built environment features that can be thought of as a point pattern – the locations of businesses, community resources or other amenities that provide goods and services that support or discourage an individual's health. To complete this objective, this dissertation adapts a suite of predominantly non-parametric Bayesian modeling techniques including Dirichlet and Gaussian Processes in order to analyze the non-linear relationships between individuals' and their environments across space and time.

In Chapter II we develop the Spatial Temporal Aggregated Predictor (STAP) model framework, to empirically determine the spatial and temporal scales at which built environment features (BEFs) have their greatest impact on human health. The framework also enables the selection of different functions that best describe the spatial-temporal exposure relationship between environment and health. This approach thus removes the unnecessary, though widely used, pre-specification of dis-

tances in time or space (e.g. 1 mile buffer) within which to measure environmental features at the population level.

Chapter III extends the work in Chapter II to identify varying forms of the spatial-temporal relationship across the population. There is a prominent interest in these effects because person and place-based characteristics shape how individuals experience and utilize the built environment. Identifying heterogeneous effects such as these hence addresses a critical question in developing place-based interventions: where and for whom are built environment interventions more likely to promote health?

Chapter IV addresses a third issue related to measurement of built environment exposures: namely characterizing the spatial distribution of BEFs around, for example, subjects' residences or places of work by identifying exposure clusters. We go on to show two ways in which these cluster assignments can then be used in a health outcomes model to identify the effect associated with the cluster assignment while still accounting for uncertainty in cluster assignment.

Chapter V presents and briefly illustrates the suite of software packages that have been developed to implement the methods discussed in the previous chapters. The `bentobox` (Built Environment Network Objects Tool Box) `R` package contains custom statistical functions, data structures and visualization functions that assist in the exploration and analysis of built environment data. We illustrate how these tools can be paired with publicly available data to more easily perform built environment analyses.

An improved understanding of the impact of environmental features on health is critically important for developing place-based strategies and policies to improve population health. This dissertation contributes to that understanding by developing methodological approaches and software tools to improve how the scientific community quantifies the influence of built environment on health.

CHAPTER I

Introduction

1.1 Overview

Motivated in part by the growth of obesity and related chronic diseases as major public health problems, there is considerable interest in understanding the mechanisms and manner through which environmental factors may contribute to chronic disease *Roux* (2003); *Garin et al.* (2014); *Renalds et al.* (2010). Observational studies have long shown links between broad neighborhood characteristics, e.g., poverty, and the development of chronic disease, and between specific environmental features and disease risk factors(*Boone-Heinonen et al.*, 2011; *Hirsch et al.*, 2014). For example, healthy food store availability and pedestrian-friendly environments can impact diet (*Boone-Heinonen et al.*, 2011) and physical activity(*Roux et al.*, 2007), and thus influence biological outcomes such as glucose levels (*Auchincloss et al.*, 2008) and obesity(*Ding and Gebel*, 2012), and ultimately increase chronic disease risk.

However, measurement challenges have limited the ability to draw causal inference from these studies and thus limited their relevance for urban design and other population-level policies. In particular, the notion of what geographic scale is relevant for measuring the impact of any given built environment feature (BEF) has posed a substantial challenge to constructing measures of exposure, as most current methods employ heuristic circular buffers around subjects' residences or workplaces. For ex-

ample, using the count of fast food restaurants (FFRs) within 1 mile of a subject's residence as the exposure metric by which to estimate association will induce bias in the – highly probable – circumstance in which all FFRs within that radius do not contribute equally to the health outcome of interest. This problem has a well documented history and is known in the literature as the Modifiable Areal Unit Problem (MAUP) (*Fotheringham and Wong*, 1991; *Guo and Bhat*, 2004; *Guo et al.*, 2011).

Similar to how the MAUP identifies that simple counts may mask the true effects of BEFs across space and time, the Uncertain Geographic Context Principle (UGCP) describes an additional limitation, in that these effects may also change across a given population or place (*Kwan*, 2018; *Macintyre et al.*, 2002). The UGCP speaks to the fact that since many urban, or increasingly suburban, populations are heterogeneous, the manner in which they interact with their environment is similarly variable. It is critical then, if science is to provide a more accurate understanding of how the built environment impacts human health and well being, to develop methods that estimate both how BEF effects may change across space and time, as well as across populations. Equally critical, is the development of software and computational methods that allow for the execution of these methods in a modern data analysis setting. The methods developed in this dissertation each provide a step towards fulfilling this higher resolution understanding. Common to all of these is the use of pairwise distances (and possibly times), between subjects and BEFs of interest. The use of distances, as opposed to counts, allows for a much more accurate estimate of a BEF's health effect.

In Chapter II, we describe the Spatial Temporal Aggregated Predictor (STAP) model, which estimates spatio-temporal functions that capture the magnitude and scales at which BEFs impact the health outcome of interest. We propose and motivate both parametric and non-parametric approaches towards estimating the spatio-temporal functions and provide software for estimating each in the `rstap` and `rsstap`

`R` packages. We apply the former parametric formulation to data from the Multi-Ethnic Atherosclerosis (MESA) Study, and identify that healthy food stores have a substantively impactful between-subject association on BMI.

Drawing from the non-parametric formulation of BEF spatial exposure discussed in Chapter II, we extend this approach to allow for the estimation of heterogeneous effects across the population. Utilizing the Dirichlet Process (DP), a popular mixture prior in the Bayesian Non-Parametric literature, this method flexibly clusters subjects with similar effects, providing investigators with an understanding of what unobserved factors may contribute to the mechanism by which BEF effects manifest. We implement this method in the `rstapDP` `R` package and apply it to census data from Los Angeles, CA in order to identify the impact of FFR exposure on student obesity. Our results identify two clusters, with one cluster's obesity risk from FFR exposure estimated to be credibly higher than the other.

Our third method described in Chapter IV builds on the idea of identifying effects across the built environment, now motivated from using a description of the built environment itself. Building on our work with the DP, we adapt an extension, the Nested Dirichlet Process (NDP), to identify clusters of BEFs' locations around subjects' residences. We also propose two methods for then using these identified cluster indicators in health outcome models, each with their own advantages and disadvantages for propagating or controlling the uncertainty in cluster classification. We apply the NDP to a dataset of California high school locations and nearby FFRs during academic year 2010 using the `bendr` `R` package. We identify six clusters of FFR distributions, one of which had a decreased risk of obesity amongst those schools consistently clustered.

Finally, as there is a large degree of overlap in the creation, manipulation and visualization of the data structures involved in the methods, we developed the `bentobox` (Built Environment Network Objects Toolbox) `R` package, described in Chapter V.

The `bentobox` package contains each of the aforementioned packages in addition to an auxiliary supporting package, `rbenvo` (Built Environment Objects in R) which facilitates the creation of the non-standard data structures commonly used by the methods in this dissertation.

By accounting for the spatio-temporal factors that determine when and where BEFs may impact health, and how these effects may vary across a population, this dissertation provides a meaningful step towards improving the scientific understanding of the built environment. Both in how it affects human health and how it may be altered so as to support health, these methods give policy makers and urban planners new tools to understand how to better shape the world in which we live.

1.2 Introduction to Point Pattern Built Environment Data

The analysis of BEF data begins with the collection and classification of the businesses and amenities near subjects into classes appropriate for the scientific question of interest. In this dissertation we use the National Establishment Time Series (NETS) data for the locations of businesses and amenities(*Walls*, 2013) for all analyses in Chapters II,III and IV. We also briefly illustrate an alternative open source data source, OpenStreetMap© in Chapter V.

The NETs database contains the locations, names and descriptions of businesses nationwide which we then use to identify particular food outlet types. A variety of methods have been proposed to ensure that business classifications are consistent and appropriate for the goods and services they offer(*Auchincloss et al.*, 2012; *Hoehner and Schootman*, 2010). In this work we used an approach that combined the standard industry code, trade name, sales volume and more, to identify supermarkets (*Kaufman et al.*, 2015).

When feasible, the data can also include the time that the subject has lived or worked at the corresponding distance from the BEF, to enable the estimation of the

temporal scale. In MESA, we calculated this as the difference in time, in years, between the study visit date and the date the business opened, or between the study visit date and the time the subject moved to the current address, whichever is shortest. We use this information to estimate the temporal scale for the relationship between HFS exposure and BMI.

The end result of the business classification and pairwise distance calculation is a "long" data frame, with multiple rows per participant, each containing the distance from each subject's residential address to each business, along with the time of exposure to that particular business. In the case of a longitudinal study with multiple visits per subject, the data would consist of multiple sets of pairwise distances and times associated with each subject's study visit (see Table 1.1).

For practical reasons, some limit will likely have to be placed on the number of businesses to include in model fitting and/or an upper limit on the distances between subject-BEF. For instance, in large, dense cities, a participant may easily accrue more than one hundred coffee shops within a five mile radius, and businesses beyond that could be excluded. This limit on the number of establishments or outermost distance, however, should be chosen on the basis of substantive reasoning, with an emphasis towards being more conservative - including more businesses - since this will have a greater chance of ensuring that subjects' exposure is estimated accurately. We describe each of these restrictions in the Chapters that follow.

subject_ID	measure_ID	BEF	Distance	Time
1	1	Fast Food	0.3	3.4
1	2	Fast Food	0.8	2.8
1	1	Fast Food	0.4	.5
1	1	Fast Food	1.3	.4

Table 1.1: Sample BEF Data Structure

1.3 Health Outcomes Data

The health outcomes data for this dissertation comes from two sources: (1) The California Department of Education (CDE) Fitnessgram project (*of Education-FitnessGram*, 2017) and (2) The Multi-Ethnic Study of Atherosclerosis (MESA) (*Bild et al.*, 2002). We describe each of these in greater detail in the Chapters where they are employed. Both datasets contain a measure of obesity, , BMI in the MESA study and an obesity classification in the CA data. These will be our outcomes of focus for this dissertation. Several subject and/or possibly neighborhood measures, such as sex, income or census tract median income are also included in the analysis to adjust for possible confounding. These are all described in greater detail in the chapters that follow.

CHAPTER II

Spatial-Temporal Aggregated Predictors

2.1 Introduction

An expanding body of research is focused on quantifying how exposure to built environment characteristics impact health (e.g., *Booth et al.*, 2005; *Charreire et al.*, 2010; *Schipperijn et al.*, 2015; *Davis and Carpenter*, 2009; *Roux et al.*, 2016; *Kaiser et al.*, 2016). The built environment refers to the human-made space in which humans live, work and recreate on a day-to-day basis (*Roof and Oleru*, 2008). As such, features of these environments constrain and/or enable everyday choices that may contribute to the development of disease. The features of the built environment are many - ranging from sidewalk availability and street connectivity to the geographic density and distribution of certain amenities such as community centers or businesses that can be mapped as point locations according to their address. In this paper we are primarily concerned with the latter and for simplicity refer to them as built environment features (BEFs). In the study that motivated this manuscript, for example, a review of different measures of BEFs used to characterize neighborhood environments was conducted, identifying that both physical and social environments are related to risk factors for cardiovascular disease (*Roux et al.*, 2016). In particular there is growing evidence to support a causal relationship between body mass index (BMI) and healthy food availability, as measured by the number of supermar-

kets, fruit-and-vegetable stores, and recreational facilities in a one mile radius around participants residential locations (*Barrientos-Gutierrez et al.*, 2017).

A key limitation in the current literature focused on estimating these effects is the unknown spatial unit within which BEFs should be measured and the functional form of the relationship between BEF effects across space. Given this lack of knowledge, studies differ in the unit used to define neighborhood environment metrics: Circular areas ("buffers") of varying radii, census tracts, block groups, and counties are all examples of different areal aggregations within which BEFs have been measured previously in the literature. Not only do these approaches make it challenging to compare and synthesize results (*Papas et al.*, 2007; *Chaix et al.*, 2005; *Leal et al.*, 2011), but they can also induce severe bias in the association of interest when the incorrect spatial unit is used (*Baek et al.*, 2016a). The latter issue has been long been recognized as the "modifiable area unit problem" (*Spielman and Yoo*, 2009; *Fotheringham and Wong*, 1991; *Openshaw*, 1996; *James et al.*, 2014; *Guo and Bhat*, 2004). An increasingly common approach is to use an individual-centered measure– for instance the availability of BEFs within a 1 mile radius of subjects' residential address, a distance equivalent to about a 20 minute walk. This approach injects subject matter expertise about the behavior of subjects (walking speed) and contextual knowledge (walking for transportation is common) to define the area for measuring BEF availability (*An and Sturm*, 2012; *Howard et al.*, 2011). However, given their *a priori* nature, these approaches fail to empirically identify the most relevant spatial unit for measurement (*Spielman and Yoo*, 2009).

Recent methodological work in this area has thus focused on furthering understanding of the distance at which these effects occur—also referred to as the spatial scale–in a data driven fashion. These scales are of relevance for decision making for both urban design and policy making, since they describe how the geographic distribution of specific amenities may support specific health outcomes. Improving

upon the buffer-based approach, (*Baek et al.*, 2016a) adapted distributed lag models (DLMs) to avoid pre-specification of the spatial scale. Instead of buffers, DLMs use counts of BEFs within a discrete series of concentric, ring-shaped areas as predictors in a regression model. Since each count is tied to the radii of the ring shaped area, the DLM thus estimates how effects of BEFs change as a function of the distance between BEFs and study participants. However, this method is limited in that it requires discretization of distance to form the ring-shaped areas, and does not enforce any substantively driven functional constraints - e.g. monotonicity of the effect with respect to distance between subjects and BEF locations. The DLM also cannot estimate spatial-temporal effects without a dramatic increase in the number of estimated parameters.

A related, although even less discussed, methodological issue is how BEF effects can vary with duration of exposure–what we will refer to as the temporal scale. That is, a longer exposure duration may be needed for a given BEF to have maximum impact. Some BEFs may have a larger temporal scale, depending on the health outcome and the pathway thorough which it confers its effect. For biological processes that take time to change, e.g., BMI, the units of the temporal scale may be in the order of months to years. On the other hand, physical activity may have a smaller time scale, potentially in the order of days or weeks, because even though it may take time to build a habit of walking for transportation, the decision to walk to a newly open gym can be instantaneous. Although there has been work on models to incorporate exposure histories(*Bandeen-Roche et al.*, 1999, 2010), we are unaware of any research that systematically investigates the estimation of temporal scales in the built environment literature.

This work introduces the Spatial Temporal Aggregated Predictor (STAP) model, which addresses questions about both the spatial and temporal scale at which BEFs contribute to health outcomes. It then demonstrates the utility of STAP through

analysis of simulated and patient-cohort data.

Section 2.2 introduces the Multi-Ethnic Study of Atherosclerosis (MESA) patient cohort which represents the subject-level data used in our motivating example *Roux et al.* (2016). Section 2.2 also presents the BEF point pattern data structure used by STAP and pertinent substantive choices that ground STAP modeling decisions. In Section 2.3 we provide a formulation of generalized linear mixed models and how the STAP framework extends this model family. Section 2.4 discusses how the STAP model is estimated in a Bayesian paradigm and how to consider prior choices and model selection. Section 2.5 demonstrates the model in the context of simulated data, while Section 2.6 showcases STAP in a real world setting via analysis of MESA data. We conclude with a discussion of future work including possible extensions to the current model in Section 2.7.

2.2 The Multi-Ethnic Study of Atherosclerosis and Healthy food store availability

2.2.1 The Multi-Ethnic Study of Atherosclerosis

The Multi-Ethnic Study of Atherosclerosis is a longitudinal cohort of men and women from six communities in the United States that seeks to understand the determinants of cardiovascular disease (CVD) prevalence, incidence and progression (*Bild et al.*, 2002). The MESA Neighborhoods Ancillary Study measured a rich set of time-varying characteristics of the participants' neighborhood environments to examine their influence on CVD, for the first five visits of MESA, spanning 2000-2010 (*Roux et al.*, 2016). These data contain participants' geocoded residential locations, Census based-characteristics, as well as the locations of a large set of amenities near study participants, among others. We focus on exposure to food stores that are considered to be supportive of a healthy diet (healthy food stores, HFS). These stores

include supermarkets, which carry a large variety of produce, among other items, as well as specialty stores including produce and fish markets. Further, we focus on the North Carolina site of MESA, given that its urban context is distinct from other sites and the dominant transportation mode (car) makes it so that the spatial scale may be larger than conventionally assumed spatial scales (e.g., 1 mile). The health outcome of interest in our analysis is body mass index (BMI), a salient risk factor for CVD.

Referencing ideas in Chapter I, we use HFS within 10 km of subject's residential address in our analysis. We view 10 km as an appropriate and conservative inclusion distance, as a similar cutoff was used in a similar analysis (*Baek et al.*, 2016b). Given the availability of data on when businesses opened and closed, which is part of the NETS database, and because we additionally have data on MESA participants' residential history (i.e., the dates when they move to a new address), we also calculated the time in years spent exposed to these nearby HFS. This time ranged from effectively 0 to twenty years at the last MESA visit.

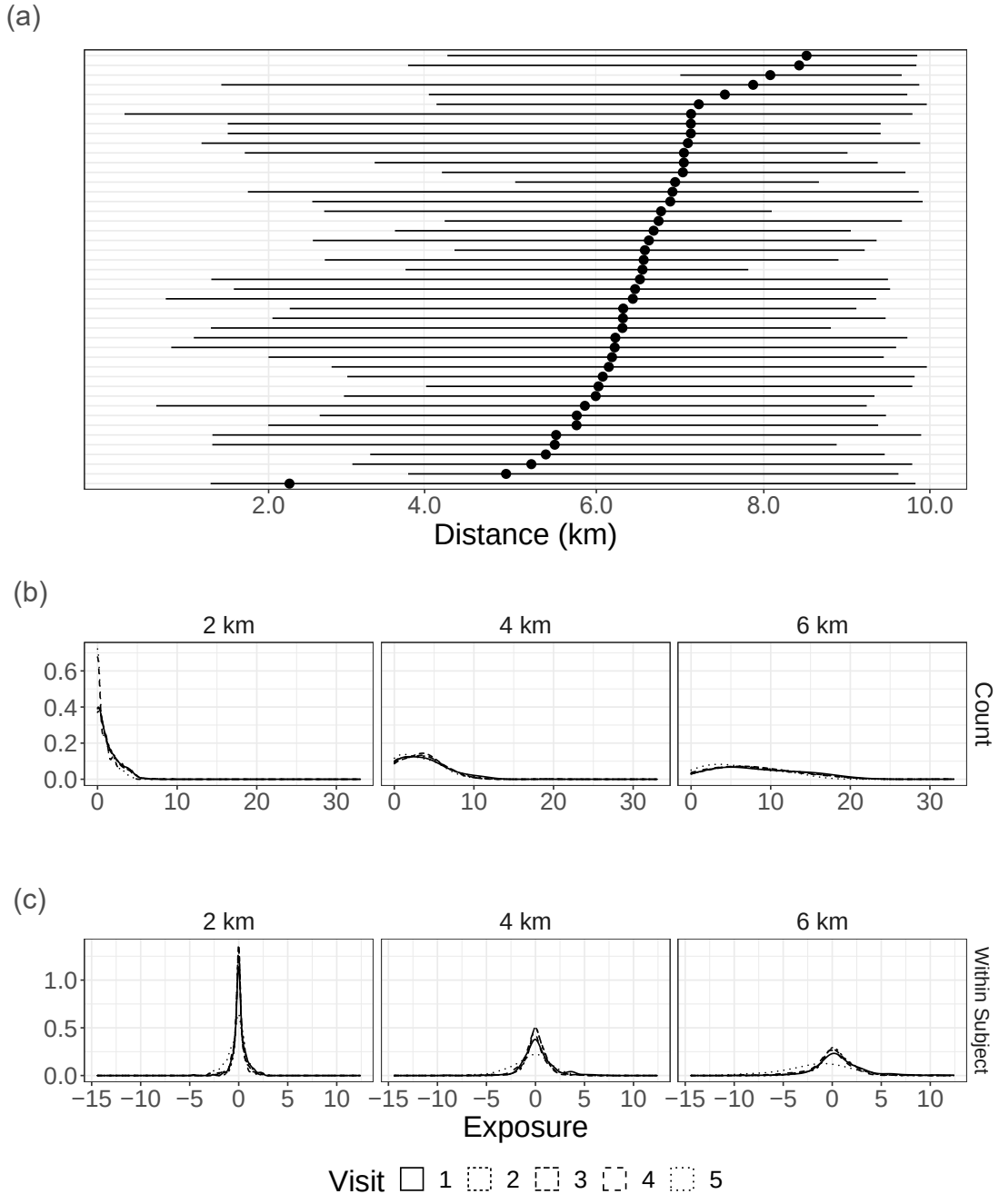
2.2.2 Descriptive analysis of distance data in MESA

Figure 2.1a shows the distribution of network distances between subjects and nearby HFS for a sample of 50 subjects at their baseline visit. For example, the first subject (first from the bottom of Figure 1a) had HFS as close as 1km, and as far as nearly 10km, but half of their HFS were within 2.5 km. This plot shows the distances that are utilized in our modeling approach described in the next section, and is informative in showing where the majority of HFS are located with respect to subjects' residence more broadly. While the first HFS is typically within 2.5 km of where subjects live (the minimum distance), the greater majority of HFS are more than 5 km away from subjects.

To further describe the data, we also use a typical exposure metric to calculate subjects' HFS exposure, namely the count of HFS within buffers of pre-determined

sizes around each subject for each visit (Figure 2.1b). These counts are a classic exposure assessment approach, and also enable us to calculate change in exposure within subject in an intuitive fashion. We calculate change within subject by subtracting their average HFS count across all visits from the subject's HFS exposure count at a specific visit. Figure 2.1c shows density estimates for the within subject change in exposure for varying buffers sizes at each visit, to give a sense of how exposure can vary across buffer sizes and visit. While there is great variability in the exposure counts across buffer size, there is relatively less variability across visits within subject. This is somewhat unsurprising given that the study period spans approximately 10 years with visits approximately 2 years apart, and, although we focused on participants who moved residences during the study period, it is likely that participants moved to neighborhoods with similar characteristics.

Figure 2.1: (a) Distribution of distances between healthy foods stores and residential locations for a sample of 50 subjects at the baseline visit (dot=median distance for each subject, lines span 2.5% and 97.5% percentiles), sorted by median distances. (b) Distribution of the number of HFS within network buffers of varying size; line types indicate different visits. (c) Within-subject differences in the BEF count within the buffers of varying size, comparing the exposure count at a given visit from the subject's average count.



2.3 The STAP Model

The STAP model extends the standard generalized linear mixed model (GLMM) regression by incorporating the estimation of the spatial and temporal scale at which one or more built environment exposures affect an outcome, as well as their corresponding coefficients.

2.3.1 The Univariate Model

For simplicity, we describe the model focusing on one type of BEF, FFRs for example, and the mean of health outcome Y_i, $\mu_i \in \mathbb{R}^1$, discussing how it can be expanded later. In order to estimate the BEF effect and spatial scale, the STAP model requires pairwise distances d between each subject and each BEF, in addition to a spatial exposure function, $\mathcal{K}_s$, defined below.

Let $\mathcal{D}_i$ be the set of aforementioned pairwise distances, $d \in \mathbb{R}^+$, between all the BEF locations and subject i. Denoting $g(\cdot), \alpha, \boldsymbol{\delta}, \boldsymbol{Z}_i$ as link function, intercept, fixed effects parameters, and ith subject's covariate vector respectively, the model is then:

$$\begin{aligned} g(\mu_i) &= \alpha + \beta X_i(\theta) + \boldsymbol{Z}_i^T \boldsymbol{\delta}, \\ X_i(\theta) &= \sum_{d \in \mathcal{D}_i} \mathcal{K}_s\big(d, \theta^s\big). \end{aligned} \tag{2.1}$$

The construction of the spatial covariate, X_i, has several implications for the interpretation of the model and permits estimation of the spatial scale that is of particular interest to built environment researchers. First, the X_i represents subject i's cumulative exposure to the particular type of BEF, accumulated according to the spatial parameter $\theta^s \in \mathbb{R}^+$ and spatial exposure function $\mathcal{K}_s$. Choosing $\mathcal{K}_s$ to be equal to 1 up until a pre-defined distance of say, 2 kilometers, would be equivalent to counting BEFs within a pre-specified buffer of radius 2 kilometers. However, we instead let

$\mathcal{K}_s$ be a non-increasing function where $K_s(0, \theta^s) = 1$ and $\lim_{d\to\infty} \mathcal{K}_s(d, \theta^s) = 0$. This corresponds to the more realistic substantive belief that a given BEFs' maximum impact is made when a subject is as close as possible to it and that the impact decreases with longer distance. To that end, we utilize the survival functions of positive continuous random variables or the complementary error function, as they satisfy all the aforementioned constraints for $\mathcal{K}_s$. While some survival functions, such as that of the Weibull distribution, may require more than one spatial parameter, i.e., $\boldsymbol{\theta}^s = (\theta_1^s, \theta_2^s)$, for now we limit our discussion to the univariate case as the extension is straightforward and demonstrated in later sections.

Second, the coefficient β is the estimated difference in $g(\mu_i)$ associated with one unit higher cumulative exposure. However, the construction of X_i allows for that "one unit" higher to be achieved by placing one new BEF at a distance 0 from the subject, all else equal. Hence β also represents the maximum effect of a single BEF on the outcome.

Finally, we define the spatial scale as the distance at which the effect of a BEF becomes negligible. This distance is thus intrinsically tied to $\mathcal{K}_s$, and can be calculated upon successful estimation of θ^s by specifying what is meant by negligible. Operationally, negligible can be defined as a proportion $p \in (0, 1)$ of the maximum effect, which occurs at distance 0 by definition of $\mathcal{K}_s$, and then find the spatial scale d^* via $\hat{d}^* = \mathcal{K}^{-1}(p, \hat{\theta}^s)$.

2.3.2 Repeated Measures Model

We now extend this framework to model temporal exposure in a setting where the ith subject has n_i visits at which the outcome is measured. Consequently we model $\boldsymbol{\mu}_i \in \mathbb{R}^{n_i}$ as a function of the previously mentioned intercept and covariates $\boldsymbol{Z}_i$, in addition to subject level effects $\boldsymbol{b}_i$ and corresponding design matrix $\boldsymbol{W}_i$. To incorporate the temporal dimension of exposure to BEFs, the exposure covariate now

aggregates over the exposure times for each subject-BEF pair, in addition to the spatial distances:

$$g(\mu_{ij}) = \alpha + \beta X_{ij}(\boldsymbol{\theta}) + \boldsymbol{Z}_i\boldsymbol{\delta} + \boldsymbol{W}_i\boldsymbol{b}_i, \tag{2.2}$$

where

$$X_{ij}(\boldsymbol{\theta}) = \sum_{(d,t)\in\mathcal{D}_{ij}} \mathcal{K}_s\big(d, \theta^s\big)\mathcal{K}_t(t, \theta^t) \quad j = 1, ..., n_i,$$

is the spatial and temporal aggregated predictor at the time of study visit j and $\mathcal{D}_{ij}$ is the set of tuple times, $t \in [0, \infty)$ and distances, $d \in [0, \infty)$ between subject i and the BEFs of interest at occasion j. The parameters for BEF of interest are $\boldsymbol{\theta} = (\theta^s, \theta^t)$ where θ^t is in the same domain as t and governs the rate at which temporal exposure to a given BEF accumulates according to $\mathcal{K}_t$. In contrast to $\mathcal{K}_s$, the temporal exposure function $\mathcal{K}_t$ is chosen so that $\mathcal{K}_t(0, \theta^t) = 0, \lim_{t\to\infty} \mathcal{K}_t(t, \theta^t) = 1$, reflecting the assumption that the maximum effect of a BEF occurs once the subject spends an infinite amount of time near the BEF and vice versus. Thus, similar to how any survival function of a positive random variable may be used as $\mathcal{K}_s$, any cumulative distribution function (cdf) of a positive random variable may be used as $\mathcal{K}_t$. Given the definitions of $\mathcal{K}_s(\cdot)$ and $\mathcal{K}_t(\cdot)$, it follows that β represents the change in $g(\boldsymbol{\mu}_i)$ when a given BEF is placed at distance 0 from the subject, for an amount of time that approaches infinity.

It is worth noting that, depending on the value of θ^t, $\mathcal{K}_t(t, \theta^t)$ will effectively evaluate to 1 sooner than in the limit to infinity. Denoting t^* as the temporal counterpart to the spatial scale discussed earlier, t^* is defined as the time at which the exposure function is $\mathcal{K}_t(t^*, \theta^t) = 1 - p$, for a small precision p. This time t^* is interpreted as the time at which a single BEF reaches its highest impact. It can be solved for in a similar manner, setting $t^* = K_t^{-1}(1 - p, \hat{\theta}^t)$.

Any application of this model to varying exposure over the differing visits k will

implicitly assume that the coefficient β of STAP exposure is equivalent within and between the ith subject. A new model formulation is required to explicitly capture these effects, a matter which we discuss next.

2.3.3 Difference in Differences Formulation

In analysis of repeated measures data, it is often desirable to estimate both the within and between subject effect associated with time-varying covariates. Decomposing these effects can avoid introducing bias into the model coefficients when the effect of within-person changes in exposure are not equal to the association of between-person difference in the exposure and the outcome. This decomposition applies to both the exposure of interest as well as adjustment for time invariant, unmeasured confounders, and can enable the interpretation of the within subject effect as a causal effect (*Neuhaus and Kalbfleisch*, 1998; *Morgan*, 2013). For example, in the built environment literature, one confounder that is seldom measured is individual's residential preference, which evolves slowly over time, if at all, and could be considered constant during the study period. While between-person differences could be counfounded by unmeasured residential preferences, i.e., due to residential selection bias, the within-person change is free of this bias. Consequently, this model specification has become increasingly popular in the built environment literature (*Hirsch et al.*, 2014).

To estimate within subject effects, β^w, STAP exposure components can be centered by the corresponding subject's mean exposure, $\bar{X}_i(\boldsymbol{\theta})$, to create the new covariate, $\Delta X_{ij}(\boldsymbol{\theta})$, which reflects the deviation from average exposure at occasion j. Estimation of the between subject effect, β^b can be accomplished by including the

latent subject specific mean exposure:

$$g(\mu_{ij}) = \alpha + \beta^b \bar{X}_i(\boldsymbol{\theta}) + \beta^w \Delta \boldsymbol{X}_{ij}(\boldsymbol{\theta}) + \boldsymbol{Z}_i\boldsymbol{\delta} + \boldsymbol{W}_i\boldsymbol{b}_i \tag{2.3}$$

$$\bar{X}_i(\boldsymbol{\theta}) = \frac{1}{n_i}\sum_{j=1}^{n_i}\sum_{(d,t)\in\mathcal{D}_{ij}} \mathcal{K}_s(\frac{d}{\theta^s})\mathcal{K}_t(\frac{t}{\theta^t}).$$

$$\Delta X_{ij}(\boldsymbol{\theta}) = \sum_{(d,t)\in\mathcal{D}_{ij}} \mathcal{K}_s(\frac{d}{\theta^s})\mathcal{K}_t(\frac{t}{\theta^t}) - \bar{X}_i(\boldsymbol{\theta})$$

Although a standard GLMM would be able to fit a model with a difference in differences parameterization using a simple pre-processing step, STAP models must be fit with the calculation of the centered covariates built-in to the overall model estimation since the constructed exposures are themselves estimated.

2.4 Estimation, Prior Choices and Model Selection

We fit the model formulated above in a Bayesian paradigm using MCMC to draw samples from the posterior distribution. We use the software in our `R` package `rstap` (*Peterson and Sanchez*, 2018) which uses the probabalistic programming language `stan` (*Team*, 2017) to implement the Hamiltonian Monte Carlo (HMC) sampler variant derived from the "No U-Turn Sampler" (*Hoffman and Gelman*, 2014). A gradient based method such as those in the HMC family is crucial for our modeling framework, as more traditional methods like Gibbs or Metropolis-Hastings Samplers are either not feasible or inefficient.

Existing literature on prior choice for standard regression coefficients and auxiliary variables applies here for all STAP parameters other than the spatial and temporal parameters, θ^s, θ^t (*Gelman and Hill*, 2007; *Gelman et al.*, 2013). Priors for spatial and temporal parameters need to be set on the basis of context-specific knowledge, including an understanding of the distribution of distances and times that will be included in the model. This is critical for ensuring model convergence as the model

may be poorly identified if one were to use, for example, an improper prior that places equivalent probability mass on all spatial scales, d^*. Indeed, while probability theory guarantees a valid posterior distribution as long as proper prior distributions are used, this does not guarantee identifiability in the Frequentist sense for our modeling approach. For example, if β in (1) above is 0, then the posterior distribution of θ^s will be equivalent to its prior distribution. Consequently, as is true of Bayesian models more generally (*Gelman et al.*, 2017; *San Martın and González*, 2010), our proposed models require forethought to be given as to how priors are placed in the context of a given data set or study location so as to arrive at a sensible posterior distribution. To further illustrate, consider a study area that could be encapsulated within a circle of 10km in diameter, with the *a priori* assumption being that the maximum plausible spatial scale that could be estimated from the model would be 10km, as distances beyond that are not plausible in that study area. Thus the prior for the spatial parameter θ_s would be defined so that plausible spatial scales, $d^* = K^{-1}(p, \theta_s)$, are at most 10km. On the other hand, if the BEF is a destination that a typical person would visit exclusively by walking, then the prior for the spatial parameter should be selected so that the corresponding spatial scales are centered at a distance that is walkable by the average person in the study sample (e.g. 1km). On the other hand, if the study area is one where the dominant mode of transportation is by car, then the spatial parameter should be selected so that the corresponding spatial scale is larger but remains bounded by the maximum distance possible within the study area. We illustrate this explicitly in our analysis of MESA data.

Given that model estimation occurs under a Bayesian paradigm via MCMC, there are a vast array of model validation and selection techniques that can be performed using, for example, posterior predictive checks (*Gelman et al.*, 2013) or the Widely Applicable Information Criterion (WAIC) (*Vehtari et al.*, 2017). The latter is demonstrated in Section 2.6.

2.5 Simulations

In the present section we demonstrate the STAP model's ability to recover parameter estimates under two broad simulation scenarios: (Section 2.5.1) differing spatial patterning of BEFs and (Section 2.5.2) differing spatial exposure functions. In the latter, we compare our model to the DLM(*Baek et al.*, 2016a). In both cases we simulate locations of subjects and BEFs within a 2 by 2 square, and outcome data under the following linear model that generates a pseudo BMI for ease of interpretation and analysis:

$$\begin{aligned} BMI_i &= 23 - 2.2Z_i + .75X_i(\theta^s = .5) + \epsilon_i, \\ \epsilon &\sim N(0, \sigma^2 = 1), \end{aligned} \tag{2.4}$$

where $X_i(\theta)$ is as specified in equation (1). In section 2.5.1 we use the Exponential exposure function to construct $X_i(\theta^s)$, and vary the spatial distribution of BEFs, thereby create different distributions of distances, d. For the simulations in Section 2.5.2, we vary the spatial exposure functions, $\mathcal{K}_s$. While we could use any combination of spatial, temporal, or spatial-temporal predictors for the purposes of demonstrating the properties of this modeling paradigm, we primarily examine spatial aggregated predictors for brevity since we found the results extend to the more complicated cases when constructing the `rstap` package(*Peterson and Sanchez*, 2018).

For each simulation replicate, the following model is estimated using the R package `rstap`(v1.1.7) on a Linux Centos 7 operating system with 2x3.0 GHz Intel Xeon Gold 6154 processors, drawing 2000 samples after burn in across 4 independent chains from the posterior:

$$
\begin{aligned}
Y_i &= \alpha + \delta Z_i + \beta X_i(\theta^s) + \epsilon_i \qquad (2.5)\\
\epsilon_i &\overset{iid}{\sim} N(0, \sigma^2)\\
\alpha &\sim N(25, 4) \qquad \delta \sim N(0, 3)\\
\beta &\sim N(0, 3) \qquad \sigma \sim C^+(0, 5)\\
\log(\theta^s) &\sim N(0, 1) \text{ or } \log(\theta^s) \sim \text{Gamma}(8, 16)
\end{aligned}
$$

Priors for covariate coefficients, $\boldsymbol{\delta}$ and $\boldsymbol{\beta}$ are set from recommendations on weakly informative priors (*Gelman et al.*, 2013). The intercept prior is set so that the true value, 23, is within three standard deviations of the prior mean, 25.

We run the simulations under two different prior settings for the spatial scale to illustrate the impact of how the prior can affect inference. The more conservative (less informative) prior for the spatial parameter θ^s is centered at a value such that the spatial exposure function $\mathcal{K}(\cdot)$ gives non-negligible weight to all but the furthest BEFs. However, this prior has a long tail so that by its 97.5th quantile, *all* BEFs affect the outcome. Equivalently, this can be understood as setting the spatial scale, d^* such that it is at the maximum possible distance between two points within the simulation square. This demonstrates how, in the absence of other prior information, a long tailed weakly informative prior can be used to estimate θ^s with good results if the effect, β, can be detected. Our more informative prior is centered at the true value of the spatial parameter, with low variability.

2.5.1 Spatial Patterning

To illustrate the robustness of STAP to differing spatial patterns in the locations of subjects and BEFs, we simulate distances, d, from settings where the locations of businesses and subjects are either independent or correlated with one another. The

correlated scenario reflects the idea that people and BEFs may be collocated near city centers, for example, as would likely be observed in real data. Specifically, locations are simulated via two-dimensional (1) Homogeneous Poisson process (HPP) and (2) Non-Homogeneous Poisson Process (NHPP). Since the number of locations can be explicitly specified under the HPP but not the NHPP, the mean of the latter was matched to the fixed number of the former. Plots showing typical simulated spatial patterns from these processes can be seen in Figure 2.2. Their corresponding distance distributions can be seen in Figure 2 in the Supplementary Material. In all cases, the exposure function is the Exponential distribution's survival function (see Figure 2.3).

Nonzero credible interval coverage, absolute difference, and model calibration are assessed by computing coverage, the percent difference between true and median parameter estimates, interval length, and Cook & Gelman Statistic (*Cook et al.*, 2006). We chose these measures since we believe precise inference, as opposed to prediction, is likely of greatest interest to investigators in this field. We provide a reference to reproducible code in the Supplementary section.

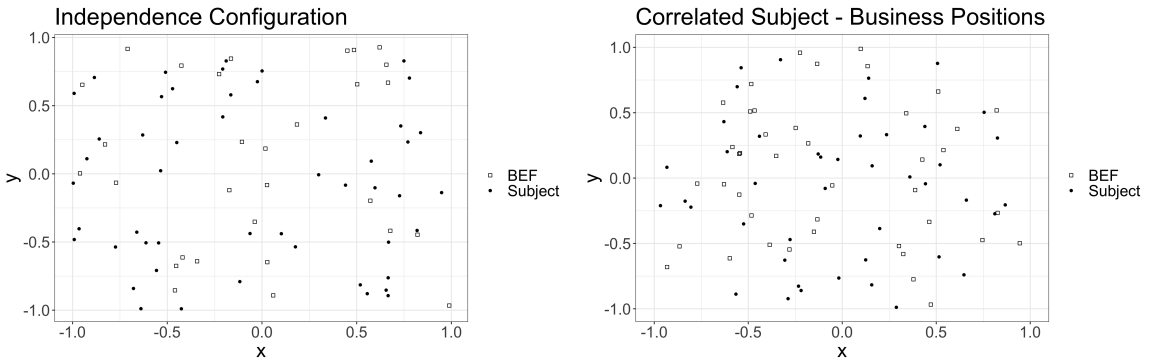


Figure 2.2: Differing Spatial Arrangement of Subjects and BEFs.

2.5.2 Spatial Exposure Functions

In order to test the explicit modeling assumption made in regard to the spatial exposure function, we simulate data using the same model as in (5) under the HPP

distance distribution using each of four different exposure functions (Figure 2.3).

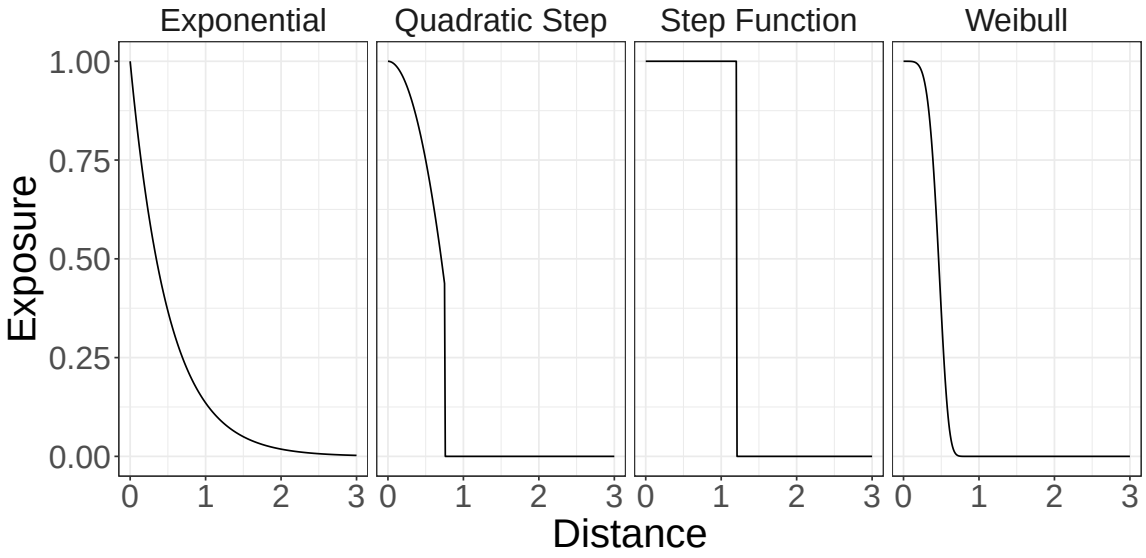


Figure 2.3: Spatial Exposure Functions.

These exposure functions are chosen to show robustness of STAPs to deviations from assumed exposure functions. Both the Exponential and Weibull distributions survival functions can be modeled explicitly as the STAP spatial exposure function, while step functions of any kind cannot because of their problematic derivatives. Similarly, we combine a quadratic decay function with a step function to create a challenge for the Exponential and Weibull having to model both the decay of the exposure effect over distance as well as a sudden end in the exposure effect.

Generating $r = 1, ..., 50$ datasets under each of these spatial exposure functions, STAP models are fit to the data using both the Exponential and Weibull spatial exposure function, using the same priors as detailed at the start of Section 2.5. From each model fit, we estimate the spatial scale, $\hat{d}^*$, using precision $p = 0.05$ for the STAP models. The DLM approach of *Baek et al.* (2016a) is used as a comparison. Briefly, this approach consists of estimating a set of smoothed coefficients $\gamma(d_\ell)$, for the model $E[Y_i] = \gamma_0 + \sum_{\ell=1}^{L} \gamma(d_\ell) X(d_{\ell-1}; d_\ell)$ where d_ℓ, $\ell = 1, ..., L$ is a discrete

grid of distances, and $X(d_{\ell-1}; d_\ell)$ is the count of BEFs within a ring-shaped area defined by concentric circles of radii $d_{\ell-1}$ and d_ℓ. We used $L = 20$, and the maximum distance $d_L = 3$. For the DLM, we estimate $\hat{d}^*$ as the smallest distance for which the credible interval of $\gamma(d_\ell)$ contains zero. For STAP models, we also calculate the estimate of the effect of the BEF at distance 0, $\hat{\beta}$ – note that there is no directly comparable counterpart to this parameter in the DLM. For each relevant model fit and generative model combination we calculate the mean percent difference from the true spatial scale and effect, e.g. $\frac{1}{50}\sum_{r=1}^{50}\frac{|\hat{d}_r^* - d^*|}{d^*}$, which are shown in Figure 2.5.

2.5.3 Simulation Results

We present the results from the informative prior, but similar plots can be seen for the conservative prior in Appendix Figures A.3 and A.4. Figure 2.4 shows the properties of STAP model estimators behave as expected in our simulated conditions. That is, as sample size increases, point estimate error and interval length decreases, while empirical coverage rates and model calibration maintain constant. In examining the differences in behavior between the two different spatial patterns, we see that the correlated subject-BEF distances regularly result in a higher bias in the estimated spatial parameter as compared to the non-correlated subject-BEF distances at a given sample size. This appears to be the only consistent difference between the two patterns, however.

Turning our attention to Figure 2.5, we see that the Weibull function performs better than the Exponential in estimating the spatial scale, d^* and exposure effect, β, in almost all cases except the simulated Exponential. This pattern is likely a consequence of the increased flexibility that comes from the Weibull's shape parameter: in estimating the step functions, this shape parameter provides the STAP model the flexibility to more closely approximate the step function's constant or near-constant exposure. However, in comparison to the Exponential model fit to the Exponential

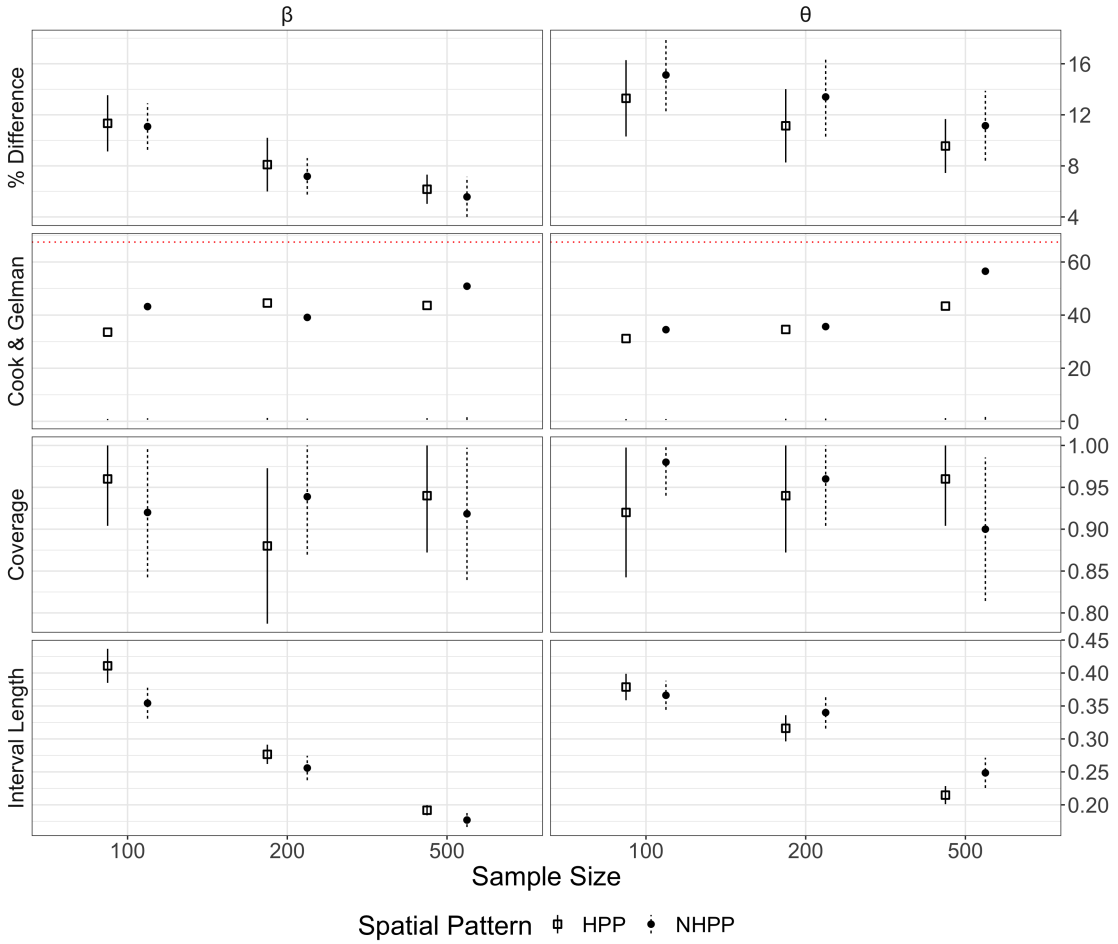
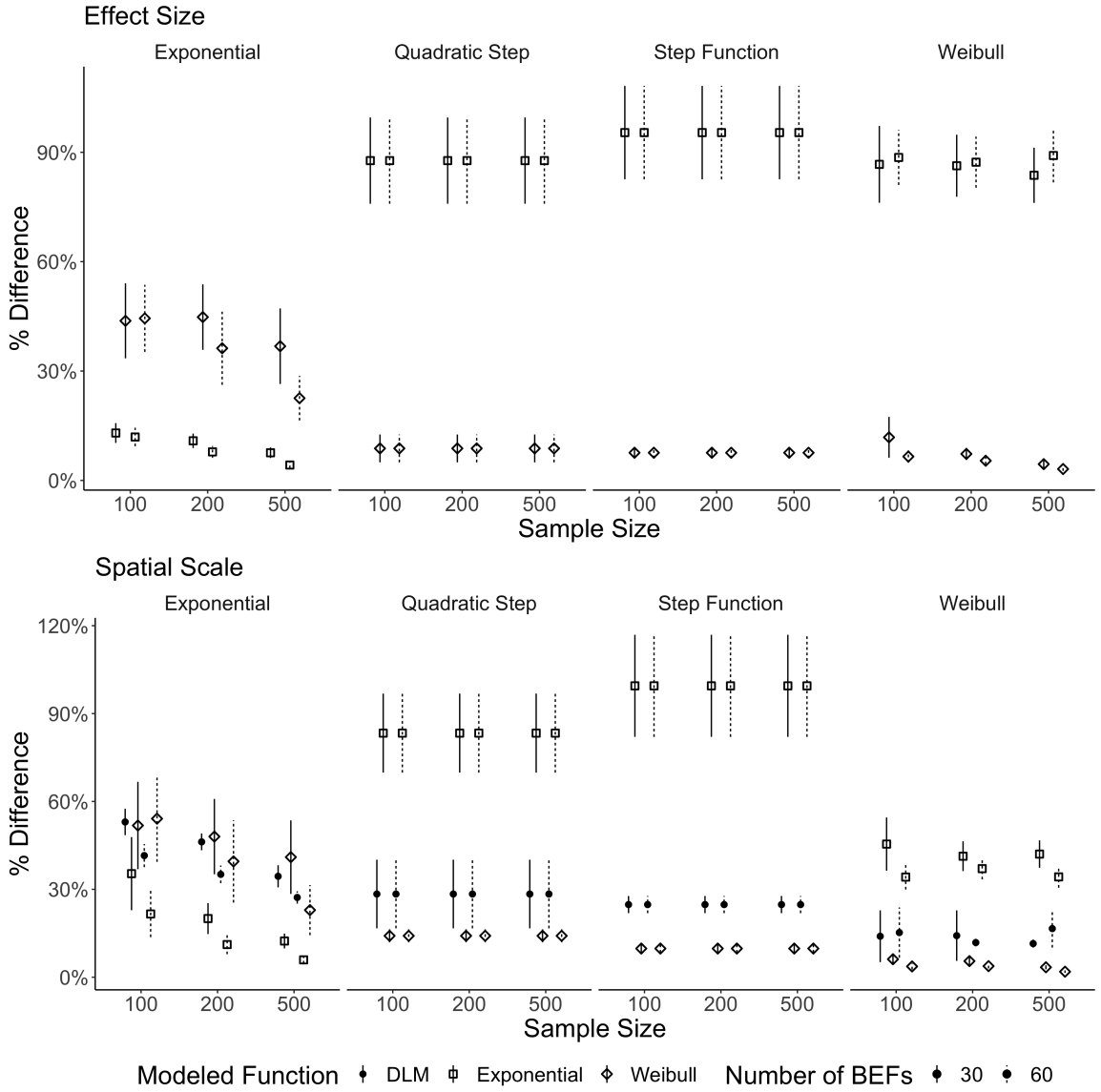


Figure 2.4: Simulation Results Evaluated by (Top Row) Absolute Difference, (2nd Row) Calibration Statistic (3rd Row) Coverage and (4th) Interval Length across Sample Size and Spatial Pattern. For all plots but Cook & Gelman, dots and intervals indicate median, 95% credible interval respectively.

Figure 2.5: Percent difference in median estimate of (Top) effect size β and (Bottom) spatial scale $d^{\star}$ from simulations varying information in sample size, generated spatial exposure function and modeled spatial exposure function.Panel title indicates the generating spatial exposure function, while dot shape indicates the median absolute difference for the modeled spatial exposure function. Line width is the 95% credible interval.



generated data, the Weibull will of course not be able to as efficiently estimate the exposure function, since the shape parameter would have to be estimated at 1 with low uncertainty in order to achieve similar results as its Exponential counterpart.

The simulation results provide guidance for investigators: using the Weibull exposure function is advantageous if there are little to no concerns regarding the ability to detect the exposure effect and there is a low level of substantive certainty regarding the functional form of spatial or temporal exposure. In contrast, if there is a greater need for estimation efficiency and a higher substantive certainty that the spatial exposure function decays exponentially, found via exploratory use of the DLM for example, then the appropriate course of action would be to use the Exponential exposure function. If there remains uncertainty about any of these conditions, model selection tools like WAIC can be used to provide evidence in favor of one exposure function over another, as we demonstrate in Section 2.6.

2.6 Relationship between exposure to healthy foods stores and BMI in MESA

This section presents the results from fitting the STAP model to data from the MESA cohort participants residing within North Carolina, to examine the relationship between availability of healthy food stores (HFS), defined in Section 2, and body mass index (BMI). As previously mentioned, we focus on participants who originally enrolled in the North Carolina site because, compared to other MESA sites, the relatively less dense urban environment at this site makes so that spatial scales could be larger than the assumed 1.6 km (1 mile) spatial scale used in prior literature. The objective is to (i) estimate the effects of HFS availability on average BMI, (ii) estimate the spatial and temporal scales at which this effect occurs and (iii) determine which exposure function better explains the spatial exposure effect of Healthy Food Stores

(HFS). We focus on individuals who moved residences at some point during follow-up, since these individuals are more likely to experience within-person change in BEF exposure and thus may provide information about the association of within-person change in BEF and within person change in BMI (i.e., β^w in Equation 2.3).

We fit the model shown below, adjusting for standard confounders including age, education, sex, and others (Supplementary Table 2):

$$\begin{aligned} E[BMI_{ij}|\boldsymbol{b}_i] = \alpha &+ \beta^w \Delta X_{supermarket,ij}(\theta^s, \theta^t) + \beta^b \bar{X}_{supermarket,ij}(\theta^s, \theta^t) \\ &+ \boldsymbol{Z}_{ij}^T \boldsymbol{\delta} \\ &+ b_{i1} + b_{i2} t_{ij}. \end{aligned} \tag{2.6}$$

We fit two models in R (*R Core Team*, 2013) using both the Weibull and Exponential Spatial Exposure functions, via the `rstap`(v1.1.06) package (v.4.0.2) on a Windows 10 operating system with a 3.7 GhZ Core i9 Intel Processor. We run four independent chains drawing 2000 samples from the posterior distribution after warming up the sampler for 2000 samples on each chain. Priors on regression coefficients were normal with mean 0 and standard deviation 3, which were selected following standard recommendations *Gelman et al.* (2013). For the spatial parameter, θ^s, we use a Log Normal prior with its scale set at 0.3, to reflect the evidence from previous work (*Baek et al.*, 2016b) that probable spatial scales are lower than 5 km. Specifically, when using the Exponential survival function to model $\mathcal{K}$, the median of the corresponding prior the spatial scale, $d^*_{.5}$ is at ≈ 4.6 km for proportion of effect $p = 0.01$. This spatial scale is in line with the estimates in the cited work. For both models we use the exponential cdf to estimate temporal exposure using the same Log Normal$(0, .3)$ prior for the temporal parameter, θ^t. As there is little previous research in estimating temporal scales, this prior is selected based on the fact that the prior encodes that belief that, with 50% probability, the majority of the effect of living near a grocery store occurs

within 4.5 years of living close to the store.

We calculate posterior credible intervals for β^w, β^b, and the spatial and temporal scales in a standard fashion, namely the median and the 2.5 and 97.5 percentiles of the posterior distributions. In addition, to help visualize the total magnitude of the effect at a given time or distance, we calculate estimates and posterior point wise credible intervals for the product of the regression coefficient (β) and the respective exposure function. First, for each posterior sample, $m = 1, ..., M$ of the spatial and temporal parameters ($\theta_s^{(m)}$ and $\theta_t^{(m)}$) we evaluate the models' spatial-temporal exposure functions along a grid of distance or time values. We then multiply the resulting quantity by the β values for the between or within effect from the same posterior iteration m. Quantiles are then calculated at each grid value, resulting in a point wise estimate of the total function effect at each point in distance or time. For instance, for distance d, the median estimate for the within subject effect across space for the Exponential exposure function would be: $\text{median}_m(\beta_m^w \times \exp(-\frac{d}{\theta_m^s}))$, where the median is taken across M total samples. The resulting functions are plotted in Figure 2.6. Note that we use the joint posterior distribution here to fully accommodate uncertainty in all model parameter estimates. However, depending on the question of interest, others may wish to marginalize or condition on the spatial or temporal parameters to arrive at an average or conditional effect estimate. Finally, we produce posterior predictive checks as a measure of model validation, provided in the supplementary material (see Figure A.5).

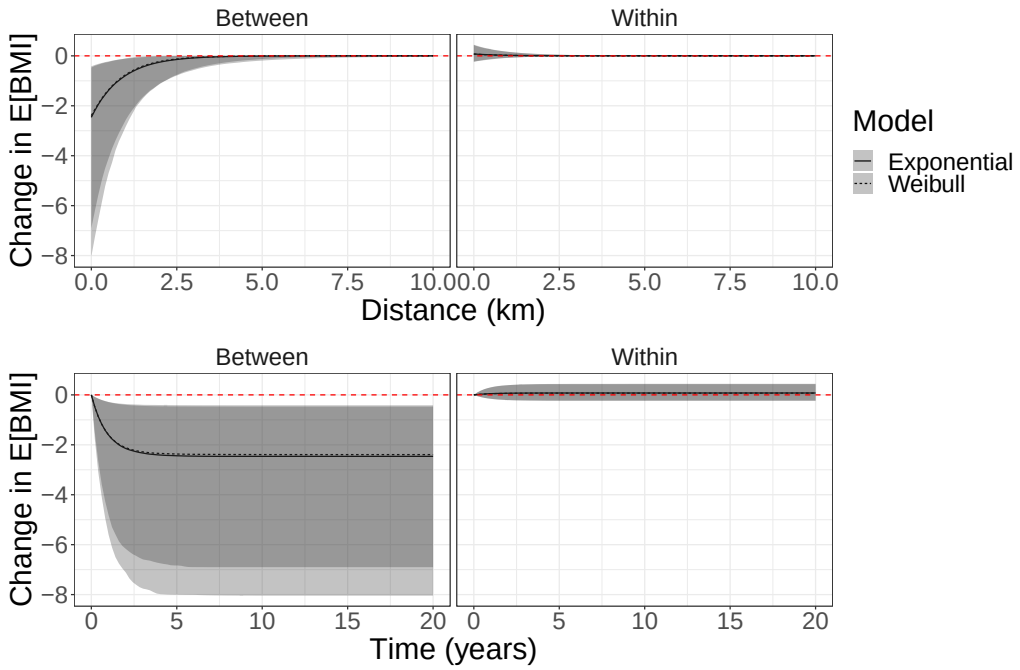


Figure 2.6: Estimated associations between BMI and healthy food store availability near North Carolina MESA participants' residential locations. (Top) Between- and within-subject associations as a function of distance from models using the Weibull and Exponential spatial exposure functions. (Bottom) Between- and within-subject associations as a function of time, using the Exponential exposure function. Shaded area corresponds to the 95% Posterior Credible Interval, interior dotted and full lines denote the median estimate of the corresponding spatial exposure function.

In the Weibull model, the median between-subject effect, β^b, was estimated to be -2.39 (95% CI: -8.04 , -.470) BMI units per unit increase in average exposure, substantially larger than the corresponding within-subject effect, β^w, estimated to be 0.07 (95% CI: -.23, .44), which is on the scale of deviations from average exposure. Similarly, the Exponential model estimates were -2.46 (-6.9 ,-.417) and 0.076 (95% CI: -.23 ,0.43) for the same parameters, respectively. This suggests that a higher average exposure to supermarkets is associated with a lower BMI after adjusting for all relevant confounders. However, there is effectively no evidence for a relationship between change in within-person level supermarket exposure and change in obesity from this model and these data.

Median estimates of the spatial scale, the distance at which the association between HFS exposure and BMI becomes practically negligible, were 4.11 km and 4.2 km in the Exponential and Weibull models, respectively. Both estimates represent a shift slightly leftward, from the median *a priori* spatial scale of 4.6 km. Median estimates for the temporal scale, the time at which the association between HFS exposure and BMI is effectively optimal , were 4.8 years with both Weibull and Exponential models showing a similar rightward shift from the prior temporal scale of 4.6 years. Note that while these results are likely sensitive to the priors used, we expect the direction of "learning", that is the direction towards these current posterior estimates to be consistent across different priors used. Alternative priors may result in "slower" or "faster" movement toward these or similar estimates depending on the uncertainty and location of the priors used to define the spatio-temporal exposure function.

In order to determine which exposure function better describes spatial exposure, we calculate the WAIC for each model fit - see Table 2.1 - and find that the Exponential model has a lower WAIC than the model fit with the Weibull spatial exposure function, suggesting the former model has better fit and greater out-of-sample predictive accuracy (*Vehtari et al.*, 2017). Intuitively, this result makes sense as the Weibull exposure function looks as if it is converging to a similar shape as the Exponential curve, only less precise due to the uncertainty associated with the Weibull's shape parameter.

	Exponential	Weibull
Spatial (km)	4.18 (2.42, 7.2)	4.11 (1.74, 10.39)
Temporal (years)	4.76 (2.67, 8.49)	4.75 (2.6, 8.37)
WAIC	27,453	103,787

Table 2.1: Estimated Spatial-Temporal Scales – Median(2.5%,97.5%) – at precision p=0.01.

2.7 Discussion

In this work we motivated, proposed, tested and demonstrated the STAP model family using both simulated data and data from the Multi-Ethnic Study of Atherosclerosis. STAP models are motivated by the need to estimate the spatial and/or temporal scale at which BEFs may effect subjects' health or health behaviors in their environment. While built environment data structures are the primary motivation of this modeling framework, other point pattern spatio-temporal data structures, such as air pollution data, could also be used in this framework.

Similar to the DLM modeling approach (*Baek et al.*, 2016a), we fully condition on distances constructed from point pattern BEF data in order to estimate a BEF's spatial scale. In contrast to their use of splines to non-parametrically estimate the spatially-varying effect, we assume a constrained functional form of spatial-temporal exposure in order to estimate a latent exposure covariate to use in regression. The use of a constrained functional form is similar to the models developed in (*Heaton and Gelfand*, 2011, 2012); however, in our approach the exposure surface (point pattern data) is fully observed and does not require modeling.

Our simulations show that little is lost by the use of parametric constraints as the Weibull exposure function is able to estimate the spatial scale of interest with less error than a competing DLM model. However, because there may be insufficient information to support the use of an exposure function as flexible as the Weibull, we've demonstrated how to use tools like WAIC *Vehtari et al.* (2017) to formally decide between differing exposure functions.

There are a number of ways in which STAPs may be extended in order to better answer more complex questions of substantive interest. For example, the spatial-temporal parameters could be constructed to incorporate a more complex hierarchical structure, providing a subject or group specific spatial-temporal scale interpretation to investigators. Additionally, incorporating interaction effects would further strengthen

the ability of the STAP family to test epidemiological questions motivating theories of how different BEFs impact groups of subjects differently. For example, the relevant spatial scale may be smaller for older adults with mobility limitations. Finally, employing computational strategies such as consensus Monte Carlo or stochastic gradient descent for these large scale data could decrease the time required to estimate parameters of interest.

Although there is a long history of Bayesian spatial-temporal models, this is, to our knowledge, the first adaptation of latent spatial variable models to fully conditioned distance or time data as a covariate in a regression model. Appropriately used, this novel methodology can increase understanding of how the community resources we live around can effect our health.

CHAPTER III

Heterogeneous Effects in the Built Environment

3.1 Introduction

The relationship between amenities in or near residential, work or school–neighborhood environments and health is receiving increasing attention, given that these environments can influence health-related behaviors and subsequent outcomes. Where spatial proximity to supermarkets is associated with diet, so too are recreational facilities associated with physical activity and fast food restaurants near schools associated with child obesity (*Baek et al.*, 2016a, 2017; *Kaufman et al.*, 2019; *Kern et al.*, 2017). Work in this area has been limited by the lack of knowledge of what geographic units are most relevant for exposure assessment, i.e. the well known modifiable unit areal problem (MAUP) (*Fotheringham and Wong*, 1991; *Spielman and Yoo*, 2009; *Wong*, 2009; *Guo and Bhat*, 2004; *Ji et al.*, 2009; *James et al.*, 2014). Additionally, there may also be measured or unmeasured person-level behaviors or characteristics that give rise to the "uncertain geographic context problem" (UGCP) (*Macintyre et al.*, 2002; *Kwan*, 2013, 2018). Whereas the former establishes that using different spatial units or spatial scales to define exposure measures will yield different estimates of association, the latter acknowledges that the most relevant spatial unit may differ from place to place or subject to subject due to place or person characteristics such as predominant transport modes in a given area.

Recent work addresses these issues by foregoing the pre-specification of the spatial unit used to construct exposure metrics. Instead, the association between proximity to amenities of interest, broadly referred to as built environment features (BEFs), and subjects' outcomes is estimated as a continuous function of distance between subjects and amenities. Whereas typical models regress the outcome on a BEF metric that depends on a pre-defined scale, these new methods use all the pair-wise distances between subjects and BEFs as inputs to the model. Specifically, in order to address the MAUP, an idealized smooth function $f(d)$ is used to represent the association between the health outcome of interest and a single BEF placed at distance $d > 0$ from the subject. Having $f(d)$ as the objective of inference enables the visualization of whether and how the association between availability of amenities and outcomes dissipates with distance, as well as estimation of the spatial scale, defined as the distance at which the association is negligible, i.e. $d : f(d) = 0$.

The function $f(d)$ has been modeled in different ways: *Peterson and Sanchez* (2018) modeled $f(d)$ parametrically, typically using exponential functions to enforce the substantive belief that the association between health outcomes and spatial availability of amenities monotonically decays across distance, e.g. $f(d) \propto \exp(-\frac{d}{\theta})$. Alternatively, *Baek et al.* (2016a) estimated $f(d)$ non-parametrically by first discretizing the distances into a grid, and using the count of distances within bins defined by the grid as predictors in a Distributed Lag Model (DLM), i.e., the count of distances within each bin are conceptualized as distributed lag predictors, indexed by the corresponding value of the grid. The coefficients corresponding to each distributed lag predictor are smoothed using splines, yielding estimates of $f(d)$ at the values of d used to construct the grid. However, the estimation of $f(d)$ at the population level, as the previous methods propose, fails to account for the concerns the UGCP raises regarding unmeasured person-level behaviors or place-level factors that may determine subject- or location- specific spatial association.

Building upon their work in DLMs, *Baek et al.* (2016b, 2017) constructed a hierarchical DLM (HDLM) allowing for the estimated $f(d)$ to vary between subjects and or locations, according to pre-specified groups (e.g., different $f(d)$ by sex), as well as unexplained variation in the association (i.e., using the idea of random coefficients to estimate $f(d)$ for individual subjects). However, the HDLM approach, has some disadvantages: (1) it uses discretized distances to estimate association across space, unnecessarily coarsening the exposure information; (2) it requires pre-specifying the groups where heterogeneity in the association may occur (covariates and or subjects); and (3) by enforcing that heterogeneity in the association estimates to occur at the subject-level through random effects, it loses possible gains in precision that could result from pooling subjects with similar levels of association.

Motivated by the desire to identify schools where pupils may be at greater risk of obesity related to the proximity of fast food restaurants (FFRs), we propose a model that clusters schools' curves, $f(d)$ according to the strength of association between the spatial proximity of nearby FFRs and child obesity. Clustering provides investigators and policymakers with a greater understanding of the kinds of relationships that exist between students and their environment as well as identifies schools where students may be at greater risk, as identifying risk groups may help prioritize population level interventions. The data for this motivating study consists of body weight status of children nested within schools across Los Angeles County during academic years 2001-2008. Distances between schools and FFRs are calculated from geocoded school addresses, supplied by the California Department of Education, and geocoded FFR business addresses from the National Establishment Time Series Database(*Walls*, 2013)

Our method uses the Dirichlet Process Mixture (DPM) prior and a spline basis function expansion to non-parametrically estimate both the number of cluster-BEF effects, and the nonlinear association functions across space, respectively. We name

our method the Spatial Aggregated Predictor - Dirichlet Process, to reflect this dual non parametric estimation, but refer to it more generally as STAP-DP, given its heritage from the previous STAP model framework and potential for modeling temporal exposure. Our approach is inspired by the work of *Rodriguez et al.* (2014) and *Ray and Mallick* (2006) on clustering functions using DPM family priors. We use the penalized spline approach developed by *O'Sullivan* (1986) and further popularized by (*Wahba*, 1990; *Wood*, 2017) to construct the estimates of the association functions, and use the DPM to cluster the spline coefficients.

Section 3.2 describes the model that estimates homo- and heterogeneous BEF effects. Section 3.3 studies the performance of the STAP-DP model in a variety of simulated data settings and discusses how the results may inform normative practice. Section 3.4 describes the application of the STAP-DP model to the motivating study on child obesity in Los Angeles. We conclude our work with a discussion of the model and future directions to explore.

3.2 Model

We now introduce the STAP-DP framework, describing how we incorporate the estimation of heterogeneous BEF effects into a regression framework. We limit our discussion to the estimation of only one BEF's effects in space, FFRs for example, as the extension to multiple BEFs is straightforward. We organize our discussion into four parts. First, we build intuition for our approach by defining the STAP estimated via spline basis functions at the population level, i.e., homogeneous effect. Then, we define how to extend the STAP model to estimate heterogeneous effects – at the latent cluster level – for a univariate outcome. In the final two sections we generalize the clustering framework for repeated outcomes measures and discuss estimation.

3.2.1 The STAP Model

Suppose a continuous outcome Y_i $(i = 1, ...N)$ and corresponding covariates $\boldsymbol{X}_i \in \mathbb{R}^{n \times p}$ are observed for a sample of N subjects. Additionally, spatial data, $\mathcal{D}_i$ which contains distances, d, between subject i and all FFRs within some substantively determined radius R, are also measured. The inferential objective is to estimate function $f(d)$, which represents the expected difference in the outcome associated with placing a single FFR at distance d after adjusting for covariates $\boldsymbol{X}_i$. Defining $F(\mathcal{D}_i) := \sum_{d \in \mathcal{D}_i} f(d)$, as the aggregated FFR effect under the assumption of additivity, we complete the initial STAP model formulation:

$$\begin{aligned} Y_i &= \boldsymbol{X}_i^T \boldsymbol{\delta} + F(\mathcal{D}_i) + \epsilon_i, \\ \epsilon_i &\overset{iid}{\sim} N(0, \sigma^2), \end{aligned} \tag{3.1}$$

where ϵ_i is the residual error, with variance σ^2.

As mentioned in Section 3.1, there are a number of approaches to model $f(d)$. In this work, we propose to model $f(d)$ as a linear combination of basis functions, $\{\phi\}_{l=1}^{L}$, which allows us to rewrite $F(\mathcal{D}_i)$ as follows:

$$F(\mathcal{D}_i) = \sum_{d \in \mathcal{D}_i} f(d) = \sum_{d \in \mathcal{D}_i} \sum_{l=1}^{L} \beta_l \phi_l(d), \tag{3.2}$$

where $\phi_l(d)$ is the evaluation of the distance through the lth basis function and β_l is the corresponding regression coefficient. In this work we use L spline basis functions defined across a set of equally spaced knots, though other knot placements or basis functions could be used. In order to avoid over fitting when L is large, the regression coefficients are regularized through the use of a quadratic penalty on $\boldsymbol{\beta}$ transformed by smoothing matrix S and tuned by penalty parameter τ. We use the difference penalty matrices of *Eilers and Marx* (1996), a widely used spline penalty formulation.

Within a Bayesian paradigm, this penalty is equivalent to specifying a multivariate normal prior with improper precision matrix τS. We adopt a variant of this Bayesian approach and, to improve computational efficiency in our more complex model formulations discussed in the next subsection, we first transform the spline basis function expansion matrix, $\Phi(d)$, such that the transformed coefficients can have independent normal priors (*Wood*, 2004, 2016). While centering constraints are often imposed on $\Phi(d)$ to avoid collinearity with the intercept in $\boldsymbol{X}$, this constraint is not needed in our model (see supplementary material). Given that r_S =rank$(S) < L$, two precision parameters for the priors are used, one for the first r_S coefficients and a second for the last $L - r_S$ coefficients:

$$\begin{aligned} &\boldsymbol{\beta}_1 \sim MVN_{r_S}\left(\mathbf{0}, \sigma^2\tau_1^{-1}\boldsymbol{I}_{r_S}\right) \quad \boldsymbol{\beta}_2 \sim MVN_{L-r_S}(\mathbf{0}, \sigma^2\tau_2^{-1}\boldsymbol{I}_{L-r_S}) \\ &\tau_z \overset{iid}{\sim} \text{Gamma}(a_\tau, b_\tau) \quad z = 1, 2. \end{aligned} \tag{3.3}$$

In (3.3) we denote $\boldsymbol{\beta}_z$, $z = 1, 2$, as the regression coefficients in the penalty range and null space, respectively. Correspondingly, τ_1 and τ_2 are the respective precisions for these separate subsets of $\boldsymbol{\beta}$. For ease of further exposition we define $\boldsymbol{\Lambda}$ as the diagonal covariance matrix which has τ_1^{-1} as the first $L - \mu$ diagonal elements and τ_2^{-1} as the last μ diagonal elements, so that the prior in (3.3) can be written simply as $\boldsymbol{\beta} \sim MVN_L(\mathbf{0}, \sigma^2\boldsymbol{\Lambda})$. We place independent conjugate Gamma priors on $\boldsymbol{\tau} = (\tau_1, \tau_2)$ so that both $\boldsymbol{\beta}$'s and $\boldsymbol{\tau}$'s conditional posterior distributions are available in closed form.

3.2.2 STAP-DP with Univariate Outcomes

In alignment with this work's goal to estimate heterogeneous effects, we replace $F(\mathcal{D}_i)$ with $F_i(\mathcal{D}_i)$ in (3.1) while allowing for clustering in the $f(d)$. Given that $f_i(d)$ is represented by the fixed spline functions and random coefficients $\boldsymbol{\beta}$, we implement

this clustering goal by placing a DP prior on the vector of regression coefficients, $\boldsymbol{\beta}$, and associated penalty parameter, $\boldsymbol{\tau}$:

$$\begin{aligned}(\boldsymbol{\beta}, \boldsymbol{\tau}) &\sim P \\ P &\sim DP(\alpha, P_0) \\ P_0 &\equiv MVN_L\left(\mathbf{0}, \sigma^2\boldsymbol{\Lambda}\right) \times \prod_{z=1}^{2} \text{Gamma}(a_\tau, b_\tau).\end{aligned} \tag{3.4}$$

In (3.4), P is a random measure drawn from Dirichlet Process $DP(\alpha, P_0)$, where $\alpha > 0$ is a concentration parameter reflecting the variability of distribution P around base measure P_0 (*Ferguson*, 1973; *Gelman et al.*, 2013). P_0 is chosen to retain the prior previously discussed in (3.3).

By placing the DP prior on $(\boldsymbol{\beta}, \boldsymbol{\tau})$, clustering is induced on the $f_i(d)$ as can be seen from the stick breaking construction of the DP: $P = \sum_{k=1}^{\infty} \pi_k \delta_{(\boldsymbol{\beta}^*, \boldsymbol{\tau}^*)}(\cdot)$. In this representation π_k represents the probability the ith observation is assigned to the kth exposure function and $\delta(\cdot)$ is the dirac-delta function. Each π_k, itself is composed of the "broken sticks" created from variables drawn from a Beta distribution: $\pi_k = v_k \prod_{u<k}(1 - v_u); v_k \sim \text{Beta}(1, \alpha)$.

Combining all these pieces together, our proposed STAP-DP model for univariate outcome Y_i takes the following form:

$$\begin{aligned}Y_i &= \boldsymbol{X}_i^T\boldsymbol{\delta} + \sum_{d \in \mathcal{D}_i} \sum_{l=1}^{L} \beta_{il}\phi_l(d) + \epsilon_i \\ \epsilon_i &\overset{iid}{\sim} N(0, \sigma^2) \\ (\boldsymbol{\beta}, \boldsymbol{\tau}) &\sim P \\ P &\sim DP(\alpha, P_0) \\ P_0 &\equiv MVN_L\left(\mathbf{0}, \sigma^2\boldsymbol{\Lambda}\right) \times \prod_{z=1}^{2} \text{Gamma}(a_\tau, b_\tau).\end{aligned} \tag{3.5}$$

A final comment is warranted regarding the choice of the number and placement of the L knots in constructing the splines. While our approach follows previous work in placing a sufficient number of knots equally across the domain of observed distances, deciding what number of knots is "sufficient" requires greater statistical judgement than in standard applications. Clusters may be more difficult to detect when the dimension on which clusters are formed (i.e., number of coefficients) is large and the between-cluster differences are small (low signal effects). Conversely, more clusters may be identified in a setting with a stronger signal and greater number of knots. Thus, L must be chosen to balance accuracy in both function estimation *and* cluster discrimination.

3.2.3 STAP-DP with Repeated Measurements

Extending (3.5) to correlated outcomes, we consider the setting in which subjects are measured repeatedly over time, for $j = 1, ..., n_i$ occasions. This results in outcome Y_{ij} $(i = 1, ..., N, j = 1, ..., n_i)$ modeled as a function of covariates $\boldsymbol{X}_{ij}$, and their corresponding coefficients $\boldsymbol{\delta}$. The distance set adopts the new visit-specific index as well, i.e., $\mathcal{D}_{ij}$, indicating it may vary over time; for instance FFRs may open and close between measurement occasions. Finally, a subset of $\boldsymbol{X}_{ij}$, $\boldsymbol{Z}_{ij}$, is included in the model, along with subject-specific coefficients $\boldsymbol{b}_i \sim MVN(0, \Sigma)$ to account for within subject variability in standard fashion (*Fitzmaurice et al.*, 2008). Augmenting (3.4)

accordingly, we arrive at our final model:

$$
\begin{aligned}
Y_{ij} &= \boldsymbol{X}_{ij}^T\boldsymbol{\delta} + \sum_{d\in\mathcal{D}_{ij}}\sum_{l=1}^{L}\beta_{il}\phi_l(d) + \boldsymbol{Z}_{ij}^T\boldsymbol{b}_i + \epsilon_i \qquad (3.6)\\
\boldsymbol{b}_i &\overset{iid}{\sim} N(\mathbf{0}, \Sigma)\\
\epsilon_i &\overset{iid}{\sim} N(0, \sigma^2)\\
(\boldsymbol{\beta}, \boldsymbol{\tau}) &\sim P\\
P &\sim DP(\alpha, P_0)\\
P_0 &\equiv MVN_L\left(\mathbf{0}, \sigma^2\boldsymbol{\Lambda}\right) \times \prod_{z=1}^{2}\text{Gamma}(a, b).
\end{aligned}
$$

3.2.4 Estimation

In order to fit models of the form described in (3.4) and (3.6), we truncate the DP so that a blocked Gibbs sampler can be used to draw samples from the posterior (*Gelman et al.*, 2013). While this sampler is fairly straightforward, it bears mentioning that $\Phi(d)$ has to be adjusted at each iteration of sampling so that any DP components associated with 0 or some small number of observations are not included in the usual matrix inversion used to estimate the mean of the conditional posterior distribution for the regression coefficients, $\boldsymbol{\beta}^* = [\boldsymbol{\delta}, \boldsymbol{\beta}]^T$. Instead, coefficients for those low-member cluster components are sampled with draws from the prior. For example, if on the mth iteration, none of the N observations are assigned to the kth DP component, then the samples of the spline regression coefficients for that iteration, $\boldsymbol{\beta}_k^{(m)}$, are drawn from a $MVN_L(\mathbf{0}, \sigma^2\boldsymbol{\Lambda}_k)$ prior, where $\boldsymbol{\Lambda}_k$ is the cluster specific covariance matrix, and the columns of zeros that would otherwise be include in $\Phi(d)$ are omitted.

We present the closed form conditional posteriors and associated algorithm in the Supplementary Material. Our algorithm is implemented in C++ which can be called from our R package `rstapDP` (*Peterson*, 2020d).

For both our simulations and applied dataset we use `rstapDP` to fit the STAP-DP

in R (v.4.0.2) *R Core Team* (2013) on a MacOS Catalina operating system with a 2.8 GhZ Quad-Core Intel Core i7 processor.

3.3 Simulations

3.3.1 Simulation Design

For a given sample size, the ability of the STAP-DP model to correctly classify subjects depends on (a) the proportion of subjects belonging to that cluster, (b) the difference in the $f_i(d)$ functional forms, and (c) the distribution of distances (i.e., exposure information) present within each cluster. As the first of these three principles follows straightforward sample-size intuitions, in this section we study the STAP-DP's ability to correctly recover cluster specific functions,$f_i(d)$, and cluster partitions in the latter two settings. Using simulated data we vary: (i) cluster effect size and (ii) distance distributions in order to see how these may impact correct cluster classification. We focus on evaluating cluster classification accuracy as it is the upstream predictor of all remaining model components, like the estimation of the $f_i(d)$, which are all standard Bayes estimators conditional on the correct cluster classification.

We evaluate our method's ability to correctly classify subjects using a partition loss function developed by (*Binder*, 1978) and used regularly in DP and other mixture model applications where label-switching may be of concern (*Lau and Green*, 2007; *Wade et al.*, 2018; *Rodriguez et al.*, 2008). Our employment of the loss function equally weights correct and incorrect classification, using the subjects' true and estimated class indicators, $\zeta_i, \hat{\zeta}_i$, respectively:

$$\psi(\boldsymbol{\zeta}, \hat{\boldsymbol{\zeta}}) = \sum_{(i,i');i<i'<N} I(\zeta_i = \zeta_{i'}, \hat{\zeta}_i \neq \hat{\zeta}_{i'}) + I(\zeta_i \neq \zeta_{i'}, \hat{\zeta}_i = \hat{\zeta}_{i'}). \tag{3.7}$$

Conceptually, (3.7) tallies the number of times that observations i and i' are incor-

rectly assigned to different clusters, when they in fact belong in the same cluster, as well as tallies when they are incorrectly assigned to the same cluster.

In each simulation setting discussed below, we generate 25 datasets and then fit the STAP-DP model shown in (3.4), truncating the DP at K= 50 and using weakly informative Gamma(1,1) priors on $\sigma^{-2}, \alpha, \tau_1$ and τ_2, respectively. We draw 2000 samples from the posterior distribution for inference via Gibbs Sampling using `rstapDP` after discarding 2000 initial samples for burn-in. Across all 25 simulations we evaluate the loss (3.7) across all $M = 2000$ iterations of the posterior samples drawn via Gibbs sampling. Given that the loss function does not have a standard range, we normalize the loss results by the maximum loss across all simulation settings, so as to make the results more interpretable relative to one another.

3.3.2 Cluster Effect Size

Our first simulation study focuses on model performance as a function of the difference between two clusters' $f(d)$, defined in (8) below, with each observation having a 50% probability of being assigned to either of these clusters. The cluster function set-up is intended to mimic a hypothetical high and low risk population scenario, in which subjects with equivalent exposure to the same BEFs experience different effects according to which risk population they belong. For each subject we generate a random number of distances uniformly so that the average number of BEFs is 15. Conditional on the number of distances, the distances themselves are then generated according to the "Skew" distribution shown in Figure 3.2. This distribution was selected in order to test our model's performance under a "worst case scenario", given that with this distribution there is relatively less exposure information at the distances where the BEF effects are non-zero. Specifically, the generative model takes

the following form:

$$
\begin{aligned}
Y_i &= 26 + .5Z_i + \sum_{d \in \mathcal{D}_i} f_{\zeta_i}(d) + \epsilon_i \qquad (3.8)\\
\epsilon_i &\sim N(0, \sigma^2 = 1) \qquad i = 1, ..., 200\\
f_1(d) &= \exp\left\{\left(\frac{-d}{.5}\right)^5\right\}\\
f_2(d) &= \nu \exp\left\{\left(\frac{-d}{.5}\right)^5\right\} \qquad \nu = (0, .25, .5, .75)\\
P(\zeta_i = 1) &= P(\zeta_i = 2) = .5;
\end{aligned}
$$

where Z_i is a covariate generated as a fair Bernoulli random variable, ζ_i is the subject specific cluster label indicating the true BEF effect, $f(d)$, for the ith observation, and ν represents the varying effect size at $d = 0$.

The relative loss as a function of the effect size ν is shown in Figure 3.1. As expected, one can see a decrease in relative loss and consequently improved classification as the effect size increases.

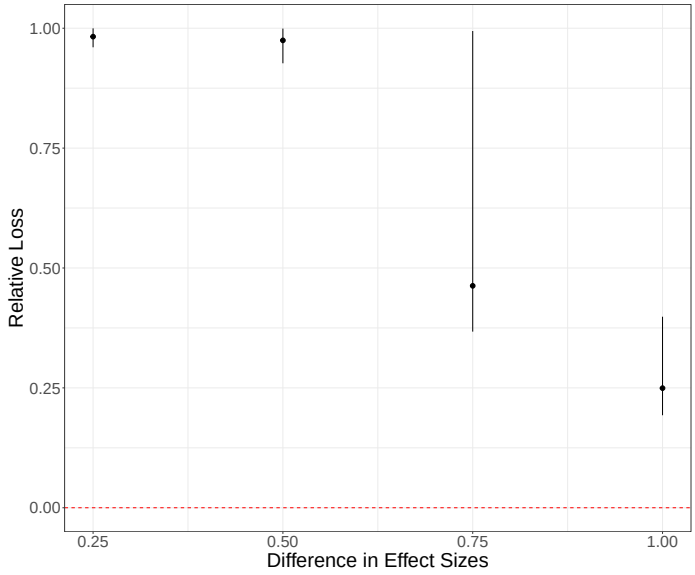


Figure 3.1: Relative loss as a function of the difference in effect size: $(1 - \nu)$; see (8) for more details. Point estimates and error lines represent median, 2.5 and 97.5 quantiles of loss across simulations, respectively.

3.3.3 Distance Distributions

As our method non-parametrically estimates cluster functions $f_i(\mathcal{D}_i)$ across continuously measured space using a basis function expansion, correct estimation of the function requires there to be BEFs observed at the relevant distances, $d : f(d) \neq 0$, within the study area of interest. Of course these "relevant" distances are not known *a priori* and so it is to the benefit of the investigator to err on the side of caution in specifying a larger study area if possible. However, despite any preparatory work that may be done to ensure an adequate area is included at the level of the sample study, it is not clear how differing distributions of distances at the latent cluster level may impact inference. For example, will suburbanites lower exposure to proximate FFRs impact the ability of the STAP-DP model to discern the impact of FFRs on their health relative to their more exposure rich urban counterparts? For this reason our second simulation study examines how exposure to different distance distributions may impact classification.

We study this problem by considering three different generative distance distributions which we label "Uniform","CA" and "Skew". The first, straightforwardly, refers to the idealistic - and unrealistic - scenario in which there is equivalent information available at all distances within the study area. The second two cases refer to more realistic situations in which there are more likely to be a higher number of BEFs found further away from the subject than close by – a consequence of area's quadratic growth as a function of distance. We create the first of these skewed distributions, "CA", by using maximum likelihood to fit a beta distribution to the distribution of distances in our motivating California data distance distribution and the second by altering a beta distribution to be a more extreme version of the first. We generate distances under each distribution for each cluster in order to examine how differing exposure patterns between clusters impact cluster classification. The densities of each of these distributions can be found in Figure 3.2.

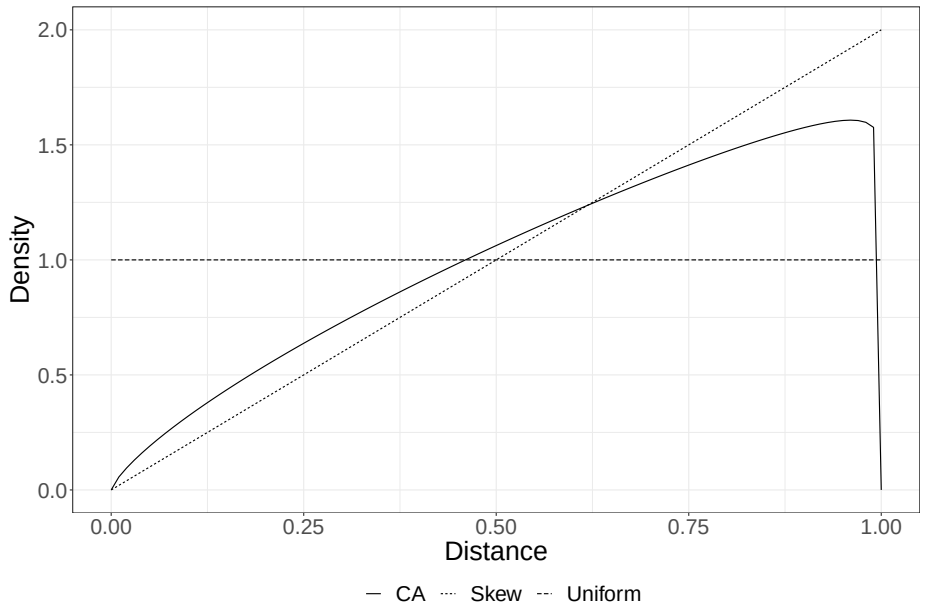


Figure 3.2: Distribution of generative distances. Line type indicates different distribution type.

Since the exposure information depends both on the distribution of distances

and the number of BEFs, we generate scenarios where the amount of information increases as a function of the number of BEFs within the same distribution. We simulate data under the same model as proposed in (3.8), with $\nu = 0.25$, to illustrate how a substantial, but not obvious, difference in cluster functions manifest across the varying distance distribution settings. Fitting our STAP-DP model under the priors and sampler settings previously described, we plot the results below in Figure 3.3.

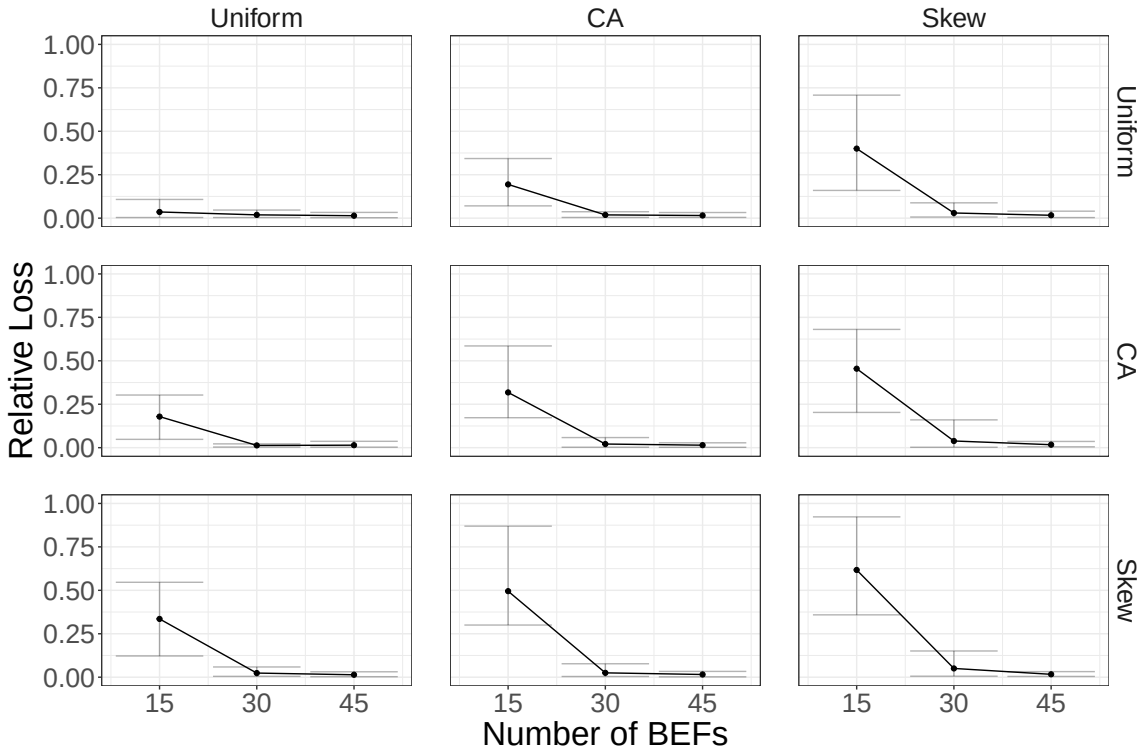


Figure 3.3: Relative loss as a function of different distance distributions. Points and lines represent median, 2.5 and 97.5 quantiles of loss, respectively. Row labels represent the distance distribution of the lower effect size cluster and columns that of the higher effect size cluster.

Figure 3.3 shows a number of patterns worth highlighting. First, across all distance distributions we observe a decrease in loss as the number of built environment features increases. This is as expected – more information or exposure results in a more easily detectable signal. Further, distance distribution combinations that include

more information result in lower levels of relative loss compared to more skewed distance distributions. There are number of cases where this can be seen in Figure 3.3, the most obvious being the top-left diagonal panel where both clusters have Uniform distance distributions; this has the lowest loss values across all panels due to the relative abundance in exposure information. This pattern holds when comparing to the more skewed distributions: The uniform-CA combination has higher error than the uniform-uniform combination when there are a relatively small number of built environment features present.

3.4 Fast food restaurants near schools and child obesity among public school students in Los Angeles

There is a pressing need to understand contextual determinants of child obesity, in order to implement population level strategies to reduce and prevent it. The food environment near schools has been proposed and studied as a contextual factor that influences children's diet, and thus obesity (*Currie et al.*, 2010; *Davis and Carpenter*, 2009; *Sánchez et al.*, 2012; *Baek et al.*, 2016a). We use data on the obesity status of children attending public schools in Los Angeles, CA, along with data on the locations of FFRs as a marker of the food environment near schools, and apply our proposed method to identify schools where children may be at higher risk of obesity, related to food environment exposures. Identifying these schools may help prioritize or tailor population-level interventions to address child obesity.

3.4.1 Data Description

Every year public schools in the State of California collect data on the fitness status of pupils in 5th, 7th and 9th grade, as part of a state mandate, using the Fitnessgram battery of tests (*Institute*, 2001). The Cooper Institute's sex-, age- and

height-specific standards for body weight are used to classify each child as "meeting the standard", "needs improvement", or "needs improvement, high risk", which correlate to normal, overweight, and obese classifications. We use the last two of these as "not meeting the standard", and use the term obesity henceforth when referring to this outcome. Fitnessgram data are available through the California Department of Education (CDE) website (https://www.cde.ca.gov/ds/). In our analysis, we use data collected during academic years 2001-2008 on 5th and 7th graders. To protect children's confidentiality, the available data consist of groups of children within the school that are defined by categorical child-level characteristics, namely sex, race ethnicity,fitness status and grade level. The percent of obese children in the group is the outcome of interest. Although we conduct our analysis at the student group level, we are able to adjust for several meaningful child-level covariates that are known to be associated with obesity avoiding the less robust, ecological analysis(*Schoenborn*, 2002).

Data on school-level characteristics are also available from the CDE website (see Table 3.1), and, importantly, so is the geocode of the school. School geocodes were used for two purposes. First, the geocodes were used to link schools to census tract level covariates. Second, the school geocodes were used to calculate the distances between the school and the geocoded location of each FFR in the LA area. FFRs were identified from the National Establishment Time Series (NETS) database (*Walls*, 2013), using a published algorithm that classifies specific food establishments as FFRs (*Auchincloss et al.*, 2012). Only FFRs within five miles of schools were kept for this analysis. This distance was chosen to be a conservative estimate as previous work estimated that the distance at which FFRs cease to have an effect on childhood obesity is approximately one mile (*Baek et al.*, 2016a).

3.4.2 Los Angeles STAP-DP Model

We fit models estimating both the population-level and latent cluster-level effects – STAP and STAP-DP models, respectively. Given the available data consisting of subgroups of children defined by the cross-classification of categorical covariates, we use the proportion of students that are obese or overweight within the subgroup as the outcome. The models are adjusting for student group and school-level covariates listed in Table 3.1. Denoting these covariates as $\boldsymbol{X}_{ijq}$ for student group $q = 1, ..., n_{ij}$, measured at year $j = 2001, ..., 2008$, attending school $i = 1, ..., N$, and using notation as described in (3.6) our model for analyzing the Los Angeles data is:

$$\begin{aligned} \%\ \text{Obese}_{ijq} &= \boldsymbol{X}_{ijq}^T\boldsymbol{\gamma} + F_i(\mathcal{D}_{ij}) + b_{i1} + b_{i2}\frac{\text{year}_{ij}}{10} + \epsilon_{ijq}, \\ \epsilon_{iqj} &\sim N\left(0, \frac{\sigma^2}{n_{iqj}}\right), \\ \boldsymbol{b}_i &\sim MVN_2(\mathbf{0}, \Sigma), \end{aligned} \tag{3.9}$$

where n_{ijq} represents the number of students in student group q during year j at school i. Given that FFRs may open or close during the study period, $\mathcal{D}_{ij}$ represents the distances between school i and FFRs available within 5 miles during year j. Similar to our simulations, we place a weakly informative Gamma(1,1) prior on each of the penalty parameters in $\boldsymbol{\tau}$, associated with each cluster regression coefficients, the residual precision σ^{-2}, and the concentration parameter α. The Gamma(1,1) prior on the concentration parameter is a common prior setting in the DP literature, reflecting the *a priori* expectation that the concentration parameter is 1, so that fewer clusters are favored (*Rodriguez et al.*, 2008; *Gelman et al.*, 2013). Additionally, we place a non-informative Jeffrey's prior on the covariance matrix for the school specific $\boldsymbol{b}_i$ vector: $p(\Sigma^{-1}) \propto |\ \Sigma\ |^{\frac{3}{2}}$. Estimation is conducted through `rstapDP`, drawing 2000 samples from each of 2 independent MCMC chains after 6000 samples have been iterated as "burn-in" on each chain. We check convergence via $\hat{R}$ diagnostic *Vehtari*

et al. (2021) and visually inspecting traceplots. We use $L = 5$ coefficients in our spline basis function expansion, and similarly use this basis to estimate the $f_i(d)$ on a grid of values, calculating the 95% point-wise credible interval at each distance grid point. We also calculate the posterior probability of co-clustering which can be arranged in a matrix $\boldsymbol{P} \in \mathbb{R}^{N \times N}$ so that $\boldsymbol{P}_{i,i'} = P($ school i is co-clustered with school i' across post burn-in iterations).

For comparative purposes, we fit a model similar to (3.9) in all ways save for restricting the $f_i(d)$ to be estimated at the population level - $f(d)$. We fit this model using Hamiltonian Monte Carlo via the `rsstap` R package (*Peterson*, 2020c), drawing 1000 samples after 1000 warm-up across 4 independent MCMC chains. Convergence is assessed via $\hat{R}$ diagnostic and we calculate the analogous posterior estimate for $f(d)$ across the same grid of distance values. For both models we produce posterior predictive checks, comparing the observed marginal outcome density to samples from the predictive distribution (see Figures A.8 A.7).

3.4.3 Los Angeles Results

Results from the STAP-DP model identify two clusters with roughly equal proportions. The $\hat{f}_i(\mathcal{D}_{ij})$ that correspond to each of these clusters and their corresponding clusters can be found in Figure 3.4 along with the population average association.

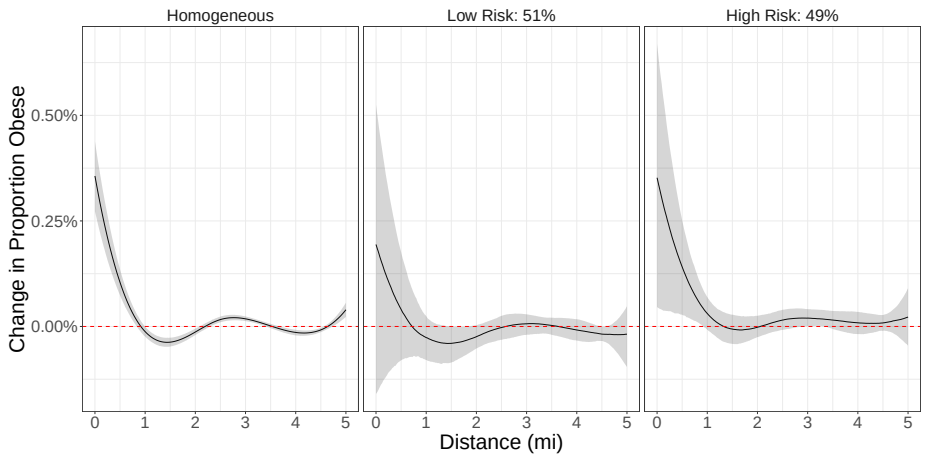


Figure 3.4: Student risk of obesity associated with FFR exposure across 5 mi.Line and band represents median and 95% posterior credible interval.

Figure 3.4 show two functions that have been labeled as High and Low Risk respectively. These names derive from the relative effect size associated with the function at and around distance 0 from the school: In the cluster labeled "High risk", one additional FFR placed at distance 0 from schools is associated with an expected 0.35% higher proportion of childhood obesity (95% CI 0.05%, .67%), all else equal. In contrast, placing one FFR at distance 0 from the schools assigned to the "Low Risk" cluster is associated with a 0.19% (-0.2%,0.5%) higher proportion of obese students. Both the muted and non-credible association are used as the justification for applying the "Low Risk" label to this latter cluster. In both clusters, the estimated associations rapidly decay with increasing distance; for the high risk cluster, the credible interval for the association first contains zero at distance 0.93 miles.

We now turn our attention to the matrix of co-clustering probabilities $\boldsymbol{P}$ which we visualize using a heat-map in Figure 3.5, after applying *Rodriguez et al.* (2008)'s hierarchical sorting algorithm to group schools with similar co-clustering probabilities together. Figure 3.5 shows that about 175 schools are consistently co-clustered within two groups, which form the core of the two clusters identified. The remaining 450 schools or so show a greater uncertainty regarding which cluster they belong to.

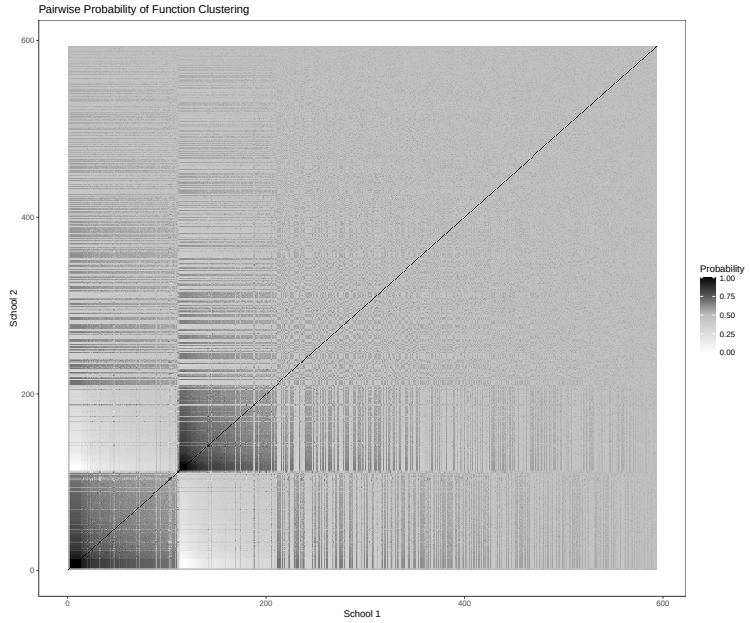


Figure 3.5: Heat map of co-clustering probabilities, that is, the probability that any two schools are assigned to the same cluster. The identity line may be interpreted as a school's probability of being clustered with itself.

This uncertainty likely stems from an insufficient number of FFRs present within the relevant ≈ 1 mile distance from schools, where the cluster effects are most discernibly different (See Figure A.6).

Further examination of the school characteristics associated with each cluster details several suggestive, though not conclusive cluster differences. These student and school-cluster specific characteristics are displayed in Table 3.1, using the cluster mode school assignment.

Differences between the two clusters of schools are fairly muted, with summary statistics across student- and school-level measures describing similar student populations and levels of wealth and education amongst the neighborhoods of schools in each cluster. The most noteworthy differences amongst the two school clusters are in the number of FFRs within 1 mile of the school, higher in the high risk group in both median and 3rd quartile compared to the low risk group. Students are more likely to be obese in the high risk group, as one would expect, compared to the low risk group.

	Overall	Low Risk	High Risk
Student Characteristics			
# Students	752,529	407,016	345,513
% Obese	51	50	53
% Female	49	49	49
Race-Ethnicity			
% Asian	3	3	2
% Black	10	11	9
% Hispanic	79	77	81
% White	8	9	8
School Characteristics[1]	N = 593	N = 299	N = 294
Total Enrollment (100's of students)	7.7 (5.0, 13.0)	7.4 (4.9, 13.4)	7.9 (5.2, 12.4)
# FFRs within 1 mile	101 (69, 143)	100 (72, 141)	102 (68, 147)
% Free or Reduced Price Meals	86 (69, 94)	85 (67, 94)	86 (69,95)
Education[2]	14 (5, 27)	13 (5, 29)	15 (6, 27)
Income[3] (1000 USD)	34 (25, 49)	34 (25, 49)	36 (25, 49)
School Type			
Elementary	468 (79%)	228 (76%)	240 (82%)
K-12	7 (1.2%)	3 (1.0%)	4 (1.4%)
Middle	90 (15%)	54 (18%)	36 (12%)
High School	12 (2.0%)	5 (1.7%)	7 (2.4%)
Other	16 (2.7%)	9 (3.0%)	7 (2.4%)
Urbanicity			
Suburban	75 (13%)	41 (14%)	34 (12%)
Urban	518 (87%)	258 (86%)	260 (88%)

Table 3.1: Characteristics of children and schools in each cluster, assigned using the mode cluster. [1] Statistics Presented: Median (IQR); N(%) [2] Median income of the households within the school's census tract. [3]Proportion of individuals with $\geq$ 16 years of education within the school's census tract.

The lack of substantial differences in measured characteristics between these two clusters is noteworthy, suggesting that – at least within this population – what would otherwise be hypothesized protective behaviors associated with education and income do not necessarily extend to school children's risk of obesity as due to proximate FFR exposure. This lack of difference in socioeconomic characteristics suggests that there are unmeasured variables that account for these differences in effects. One possibility could be the type of FFRs proximal to the schools, e.g. chains vs non-chain FFR's. Alternatively, the differences could reflect different quantities of FFR's around these schools, as the assumption of FFR exposure additivity may be more justified in one subgroup than in another, resulting in different effects as compromise.

3.5 Discussion

This work proposed a modeling approach to identify heterogeneity in distance-dependent BEF effects. By allowing flexibility both across space and identifying subgroups of subjects with different effects, this modeling framework addresses two problems raised in the built environment literature, namely the MAUP and the UGCP, respectively. The modeling approach was shown to work well in both simulated data, as well as the data that motivated this work, concerning children's obesity status and proximity to FFRs near their schools. While spatial point pattern built environment data are the primary motivation for this methodology, it could be also be applied to temporal or spatio-temporal data, the latter which we discuss in greater detail below.

Similar to the HDLM proposed by *Baek et al.* (2016b), we seek to allow for differences across subjects, or other substantively defined groups like schools, in the BEF associations across space. In contrast to that work, we pool subjects with similar association effects through the DPM, allowing us to identify latent risk subject groups.

In simulations our model demonstrated classification robustness to differing dis-

tributions of distances and expected improvement in classification due to increased information through BEF exposure or effect size. Our analysis of Fitnessgram data illustrated how one can analyze these data in terms of the spatial effects estimated as well as the characteristics associated with each latent cluster. The software to fit this model and perform the necessary auxiliary functions is freely available through our R package `rstapDP` (*Peterson*, 2020d).

There are a number of future directions with which to take this work. One obvious direction would be to extend the modeling framework for more general exponential error distributions, though this makes estimation more difficult, as the posterior distribution of $\boldsymbol{\beta}$ is no longer available in closed form. Work by *Ferrari* (2020) has used a Riemann Hamiltonian Monte Carlo sampler in this context for models similar to ours, without smooth functional terms. This could provide one avenue to pursue. Another direction to explore would be to incorporate temporally indexed BEF data to enable spatio-temporal function estimation via tensor product of the spline basis function expansion used here. This approach would allow for cluster estimates across space and time, increasing the dimensionality and consequently, relevancy, of this work to more precisely target and understand how environments shape health and health behaviors across both time and space.

More general future work could examine the role of the concentration parameter in determining the number of clusters. In our work we use a standard prior in the literature, but more informative priors lead to different a number of clusters and cluster functions (See Table A.2). This sensitivity may be seen as a limitation and previous work discusses these issues at length (*Ishwaran and James*, 2001; *Miller and Harrison*, 2014; *Miller*, 2014; *Miller and Harrison*, 2018; *Rodriguez et al.*, 2014; *Liu et al.*, 2018).

Finally, while there have been numerous methods to identify associations between subjects and BEF exposure we believe this to be the first to utilize techniques in both

the Bayesian and functional non-parametric literature to identify heterogeneous BEF effects across a population.

CHAPTER IV

Identifying Health Relevant Built Environment Patterns

4.1 Introduction

The dramatic increase in child obesity is one of the most pressing public health issues of the 21st century (*Sacks et al.*, 2012). The potential causes of lack of energy balance that result in child obesity have been widely studied, and the need for population-level interventions, beyond individual-level treatments has been strongly emphasized by the research community and policy makers alike (*McGuire*, 2012). Place-based interventions are one realm of population level approaches that seek to modify neighborhood environments in ways that can support residents' health promoting behaviors. Within this type of approach, changes to the distribution of health-supportive (or detrimental) amenities within neighborhood environments have emerged as a possibility, given that the built environment – the human made space in which humans live, work and recreate on a day-to-day basis – constrains everyday health-relevant choices (*Roof and Oleru*, 2008).

In particular, the potential contribution of the food environment near schools (e.g., fast food restaurant availability) to child obesity has been studied extensively (*Currie et al.*, 2010; *Davis and Carpenter*, 2009; *Sánchez et al.*, 2012; *Baek et al.*,

2016a), given that children spend large proportions of their waking hours and consume a large proportion of their food within and near the school environment. While the body of evidence supports these connections broadly, different approaches to conceptualize exposure make it challenging to more fully understand the health effects of environmental exposures, as well as identify where interventions may be especially needed. In order to assist policy makers with these challenges, methods need to be developed that both (i) identify different spatial patterns of exposure and (ii) link these patterns to health outcomes quantitatively. Exposure patterns, compared to continuous exposure measures, may make it more straightforward to identify places in higher need of interventions.

Previous work in environmental epidemiology has approached these problems by first clustering some measure of built environment features (BEFs), e.g. the number of fast food restaurants (FFR) within a mile, and then incorporating cluster assignments as a categorical predictor into a second stage regression model. For example, *Wall et al.* (2012) used a spatial latent class analysis (LCA) to cluster multivariate measures of the built environment, including the density of food outlets within 1 mile of the subjects residential location, and subsequently estimated the association between cluster membership and adolescent obesity. Like *Wall et al.* (2012)'s analysis, it is common to use a simple count of BEFs within some pre-specified buffer (e.g. 1 mile) as an exposure measure (e.g., *An and Sturm* (2012); *Howard et al.* (2011)). However, clusters identified with these traditional exposure summaries ignore the spatial distribution of amenities within the buffer. This spatial distribution is relevant from a mechanistic perspective because BEFs closer to schools are easier to access, as well as policy relevant since the distribution could inform built environment interventions such as zoning laws to curtail exposure. Finally, plugging in an estimate of cluster membership as a predictor in a health outcome model does not account for the uncertainty in the estimated cluster assignment label, leading to potentially

incorrect inference of the associated health effect.

Motivated by the need to better understand the association between exposures to FFRs near schools and child obesity, this paper has two complementary goals. First, we aim to develop a clustering procedure that provides interpretable groupings of BEF exposure while taking into account the spatial distribution of BEFs. For this goal, we work with the geographical coordinates of BEFs and schools, modeling the set of distances of each school to its nearby BEFs as a realization of a 1-dimensional point pattern process with a school-specific intensity function. Clusters of schools are formed by clustering the intensity functions of these point patterns using a Nested Dirichlet Process (NDP). Working with point-level data and using the actual distances between schools and BEFs, rather than aggregating the data at the areal-unit level, allows us to maintain the level of granularity needed to investigate the effect of the spatial distribution of BEFs around schools on children's obesity. In particular, our approach to deriving the schools' cluster assignments is based on the distribution of distances of schools and their surrounding FFRs, but not the quantity of FFRs. This clustering approach enables us to separate the contribution to obesity associated with the number of FFRs near schools from the association of obesity with the relative proximity of FFRs to the school, thereby providing new insights compared to prior work. Second, we show two ways to use the output from the NDP clustering model to address cluster assignment uncertainty when using clusters as predictors in a regression that evaluates the association between FFR exposure near schools and obesity risk of students in those schools.

As a statistical genre, clustering methods vary widely, from the model-based finite mixture models (FMM) (*Diebolt and Robert*, 1994) and previously mentioned LCA (*Wall and Liu*, 2009), to the more algorithmic k-means style methods (*Hartigan*, 1975; *Friedman et al.*, 2001). Each of these have varying strengths and weaknesses according to the problem at hand. Notably, FMMs, K-means and LCA share the

important assumption of pre-specifying the number of clusters that should be found in the data. LCA and K-means also make parametric assumptions about the relevant distribution or metric, respectively, that should define the clusters. In our pursuit of examining the contribution of both the spatial distribution and quantity of FFRs around schools on obesity without strong parametric assumptions or pre-specification of the number of clusters we employ a NDP approach. We use the NDP to flexibly cluster schools according to the spatial distribution of BEFs around them, which we assume is regulated for each school by the intensity function of an Inhomogenous Poisson Process (IPP). The NDP allows us to estimate these functions without enforcing any strong parametric constraints on the shape of the intensity function or the number of clusters. Akin to *Xiao et al.* (2015), our model for the intensity function of the IPP factorizes the intensity function into the product of a normalized intensity function modeled non-parametrically, and a total intensity. Nevertheless, our model is different from that of *Xiao et al.* (2015) in multiple ways. First, we are modeling an IPP over space with the goal of identifying common patterns in the spatial distribution of FFRs around schools, while the latter uses a marked IPP over time in order to examine the inter- and intra-annual variation of hurricanes frequency. Additionally, while *Xiao et al.* (2015) invokes a dependent Dirichlet process (DDP) (*MacEachern and Shen*, 1999; *MacEachern*, 2000) to capture the temporal dependencies across years in hurricanes' frequencies when modeling the intensity function, our model employs a Nested Dirichlet Process to identify clusters of schools with similar spatial distributions of FFRs near them.

Additionally, in accordance with our conceptual objective (i), the NDP provides cluster assignment labels which can be processed and used in a second-stage regression analysis to estimate associations between the BEF's spatial distribution and a health outcome of interest. Second-stage models raise the need to accommodate uncertainty in estimated exposures (in this case cluster assignment), a need that has been the topic

of several papers (*Chiang et al.*, 2017; *Graziani et al.*, 2015; *Wall et al.*, 2012; *Wade et al.*, 2018). Here, we explore two approaches to using the output of our clustering model in a second-stage analysis as a way to deal with the challenges of making cluster assignments; namely the NDP yields a varying number of cluster assignments in the posterior samples. One approach relies on using a conservative "consensus of cluster assignments" as determined from cluster-specific uncertainty bounds (*Wade et al.*, 2018). The second approach avoids the use of a single cluster assignment by using each school's vector of co-clustering probabilities with other schools as a measure of multivariate exposures, inserting the vector as a predictor into a health outcome model through a Bayesian kernel machine (BKMR) regression approach (*Bobb et al.*, 2015; *Valeri et al.*, 2017; *Coker et al.*, 2018; *Wang et al.*, 2018). This latter approach is an innovation in terms of expanding the applications of BKMR, as well as a way to utilize a clustering model's output to address classification uncertainty.

The layout of the paper is as follows: Section 4.2 discusses the data sources used in our analysis of child obesity in relation to FFR occurrence near their schools, namely data on children's obesity status and school characteristics obtained from the California Department of Education, and food outlet locations from the National Establishment Time Series (NETS) database. The section includes discussion of some nuances involved in handling this and similar data for our proposed statistical methodology, as well as some preliminary data analysis. Section 4.3 describes the modeling approach that we propose for clustering schools with respect to the spatial distribution of FFRs around the school and the two modeling frameworks for the second stage health analysis. Section 4.4 contains the results from fitting our models to the California data. We finish with a discussion of the contribution our work makes to the built environment literature, limitations of our approach and possible methodological extensions.

4.2 Data on child obesity and food environment near schools in California

4.2.1 Data sources and study sample

Each spring, public schools in the State of California collect data on the fitness status of pupils in 5th, 7th and 9th grade, as part of a state mandate, using the Fitnessgram battery. The Cooper Institute's sex-, age- and height-specific standards for body weight are used to classify each child as "meeting the standard", "needs improvement", or "needs improvement, high risk", which correlate to normal, overweight, and obese classifications. We use the last two of these as "not meeting the standard", and use the term obesity henceforth when referring to this outcome. Fitnessgram data are available through the California Department of Education (CDE) website (https://www.cde.ca.gov/ds/). In our analysis, we use data collected during academic year 2009-2010 on 9th graders only, since high school youth are more likely to be exposed to the food environment surrounding their school (e.g., students may leave the campus for lunch).

Data on school-level characteristics are also available from the CDE website (see Table 4.1), and, importantly, so is the geocode of the school. School geocodes were used for two purposes. First, the geocodes were used to link schools to census tract level covariates. Second, the school geocodes were used to calculate the distances between the school and the geocoded location of each FFR in California. FFRs were identified from the National Establishment Time Series (NETS) database (*Walls*, 2013), using a published algorithm that classifies specific food establishments as FFRs (*Auchincloss et al.*, 2012). Only distances shorter than one mile were kept for this analysis. This distance was chosen on the basis of previous work that estimated that the distance at which FFRs cease to have an effect on childhood obesity is approximately one mile (*Baek et al.*, 2016a). Finally, we calculated the distance

between all schools, to derive a data set of schools so that there are never two schools within one-mile of one another, in order to satisfy independence assumptions used in the analysis.

4.2.2 Preliminary analysis

The dataset is comprised of 420,085 children who attended 1,193 high-schools. Across all high schools, 40% of the 9th graders were observed to be obese. Of the high-schools, 767 had at least one FFR within one mile and 426 had zero. Although the second stage analysis for the health outcome includes *all* schools in the dataset, regardless of whether they have FFRs or not within a mile, the first stage analysis that derives clusters of school with similar spatial distribution of surrounding FFRs excludes the 426 schools that do not have any FFRs within one mile of their location.

Descriptive statistics of the schools are presented in Table 4.1, for the entire dataset and for the two subsets of schools without FFRs or with at least one FFR within a mile. As the Table shows, aside from having at least one FFR, schools included in the first stage analysis are generally more likely to be located in urban areas (46%) compared to schools not included (27%). Schools included in the first stage analysis varied both in terms of the number of nearby FFRs and in terms of their spatial distribution, both important aspects of BEF exposure. Among these schools, 45% had 1 to 4 FFRs within a 1-mile buffer while the rest of the schools had at least 5; additionally the median (Q1-Q3) distance to the first FFR was 0.4 (0.3-0.6) miles.

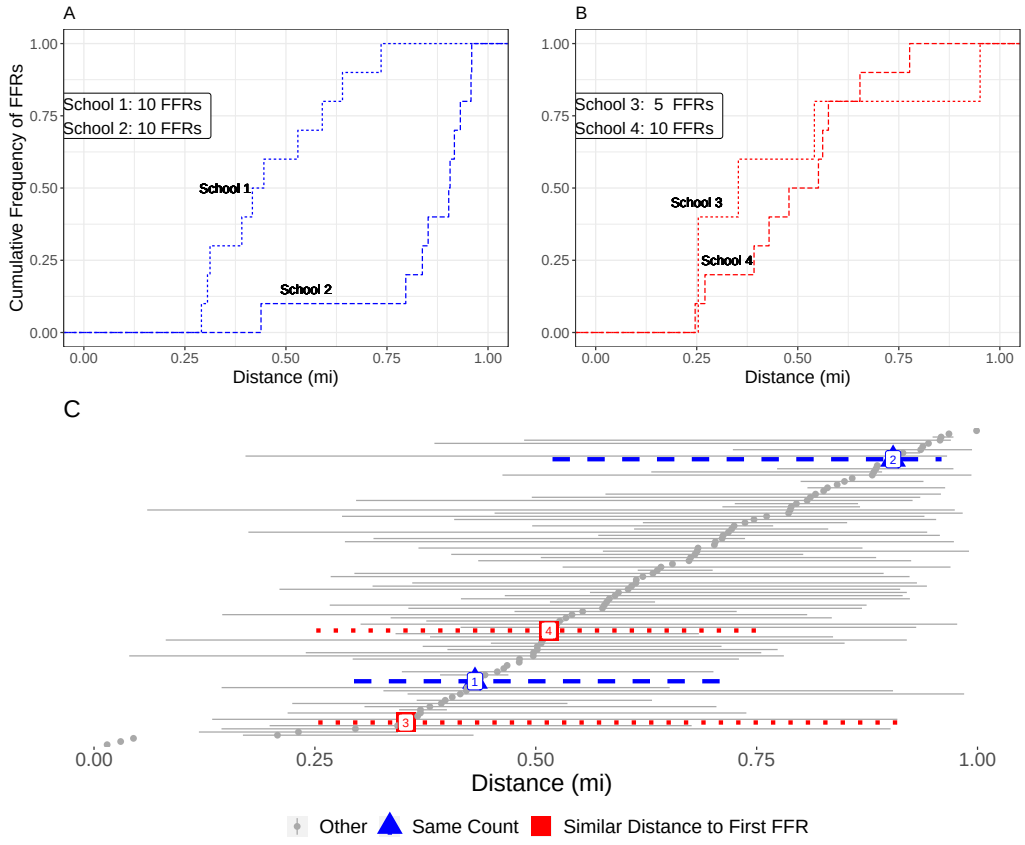
A richer understanding of schoolchildren's exposure to FFRs can be gleaned by examining the empirical cumulative distribution function (ECDF) of the distances between each school and its neighboring FFRs. The ECDFs for four schools are shown in the top panels of Figure 4.1. These two panels illustrate how traditional measures of built environment exposure, using simple counts or distance to the closest

FFR, may fail to incorporate meaningful aspects of spatial exposure.

	Subset of schools with		All schools
	≥ 1 FFR nearby	no FFRs nearby	
Children			
Number	298,903	121,182	420,085
% Obese	40%	40%	40%
Schools			
Number	767*	426	1,193
Income[1] ($1,000 USD)	59.1 (43.4-78.6)	52.3 (40.2-74.4)	56.7 (42.5-77.1)
Education[2]	23.7 (12.6-38.6)	19.6 (11.2-32.6)	22.2 (12.1-36.2)
Closest FFR (Miles)	0.4 (0.3-0.6)	-	0.4 (0.3-0.6)
FFR Quantity			
0	0	100	36
[1,4]	45	0	29
$\geq$5	55	0	36
Urbanicity			
Rural	10	39	21
Sub-Urban	44	34	40
Urban	46	27	39
Majority Race			
African American	1	2	1
Asian	4	2	4
Hispanic	29	27	28
No Majority	14	9	12
White	53	59	55

Table 4.1: Descriptive statistics for children and schools in the analytic dataset. Summary statistics for continuous variables are Median (Q1-Q3) and column percentages for categorical variables. [1] Median income for households in the school's census track. [2] Proportion of individuals with $\geq$ 16 years of education. *17 schools have missing data on obesity.

Figure 4.1: Panel A: Distribution of distances from the school to nearby FFRs for two schools with 10 fast food restaurants (FFRs) within a 1 mile radius. Panel B: Distribution of distances from the school to nearby FFRs for two schools that have the same distance to the closest FFR. Panel C: distribution of distances to FFRs for a sample of 100 schools. For each school the plot shows the range of distances between the 2.5th and the 97.5th percentile. Schools are sorted by median distance to FFR. Darker dashed and dotted lines represent the four schools depicted in panels A and B of this figure.



Specifically, Figure 4.1A illustrates how schools that may have a similar number of FFRs within a given distance may be characterized by a dramatically different spatial distributions of the surrounding FFRs. Likewise, Figure 4.1B illustrates that while certain schools may have similar distribution of distances to FFRs, the total number of nearby FFRs might be completely different. Characterizing exposure to FFRs on the basis of these traditional measures or clustering schools based only on how similar these two exposure metrics are across schools could miss meaningfully important

aspects of spatial exposure, and consequently not fully capture the effect of exposure to unhealthy food on health. Figure 4.1C illustrates the distribution of school-FFR distances for 100 randomly selected high schools, including the aforementioned four illustrative schools, further demonstrating the wide variability in exposure to FFRs in our dataset. In order to accommodate the limitations in capturing exposure to FFRs discussed above, our modeling approach considers *all* the distances from each school to its nearby FFRs, modeling this set of distances as a realization of a 1-dimensional point pattern process whose school-specific intensity function is estimated non-parametrically.

4.3 Model

As described in Section 4.1, our analysis of the effect of exposure to FFRs on obesity is based on a two-stage approach: (i) a first stage model that characterizes the main patterns of school-level exposure to FFR, by deriving characteristic profiles of the spatial distribution of FFRs near schools; (ii) a second stage model that uses the output from stage 1 in a regression model to examine the association between patterns of exposure with child obesity. In the following sections we provide details of each of these modeling strategies, including specific aspects of our models' implementations and estimation.

4.3.1 Clustering model

To characterize the food environment near schools, our clustering approach focuses on the point processes describing the relative locations of the FFRs in the immediate vicinity of the schools, rather than the global 2-dimensional point process representing the location of FFRs across the entire state of California. Specifically, let r_{ij} be the distance between the ith school ($i = 1, ..., N$) and the jth nearby FFR ($j = 1, ..., n_i$): each r_{ij} belongs to the interval $(0, R) \subset \mathbb{R}$, with maximum distance R chosen on

substantive grounds. Since the schools in the sample are relatively far from each other (by at least R), the distribution of distances for one school does not inform on the distribution of distances for another. Thus, for each school i, we model the random subset $\mathcal{D}_i = \{r_{ij}; j = 1, ..., n_i\}$ as a realization from a one-dimensional Inhomogeneous Poisson Process (IPP) with intensity function $\lambda_i(r), r \in (0, R)$. We further decompose the intensity function $\lambda_i(r)$ as $\lambda_i(r) = \gamma_i f_i(r)$ with γ_i representing the expected number of FFRs within radius R from school i and $f_i(r)$ denoting a normalized density. Thus, the ith school's contribution to the likelihood is:

$$p\left(\mathcal{D}_i|\gamma_i, f_i(r)\right) \propto \gamma_i^{n_i} \exp\{-\gamma_i\} \prod_{j=1}^{n_i} f_i(r_{ij}). \tag{4.1}$$

Assuming independence between schools, the full likelihood is obtained by taking the product over the N schools' likelihood contributions. In (4.1) we note that the likelihood is separated into a component that handles the number, n_i, of FFRs for each school i, and a component that, given n_i FFRs surrounding school i, evaluates the density at each of the n_i distances. For our purposes, $\gamma_i, i = 1, ..., N$ are considered nuisance parameters that do not affect the estimation or interpretation of the $f_i(r)$ beyond what has been previously discussed. In the health outcome model we use the observed n_i's directly as a predictor, instead of their expected values γ_i's, in accordance with our aim to differentiate between the separate effects on childhood obesity of the observed FFR quantity and the FFRs' spatial distribution.

Our goal is to simultaneously model and cluster the FFR spatial density functions $f_i(r)$, $i = 1, \ldots, N$, in a non-parametric fashion. The non-parametric estimation of a single $f_i(r)$ could be accomplished by using a Dirichlet Process (DP) mixture model (*Gelman et al.*, 2013). However, in order to fulfill our goal of non-parametrically estimating and clustering the $f_i(r)$'s themselves, we use a NDP modeling approach. The NDP enables us to achieve both of these objectives through the use of two DP's

simultaneously. Specifically, we express each $f_i(r)$ as:

$$
\begin{aligned}
f_i(r) &= \int \mathcal{K}(r|\boldsymbol{\theta}) dG_i(\boldsymbol{\theta}) \\
G_i &\overset{iid}{\sim} Q \\
Q &\sim DP(\alpha, DP(\rho, G_0)),
\end{aligned}
\tag{4.2}
$$

where, $\mathcal{K}(r|\boldsymbol{\theta})$ is a mixing kernel with parameter vector $\boldsymbol{\theta}$, and the distribution G_i is drawn from the random distribution Q on which we place a NDP prior. In (4.2), $DP(\rho, G_0)$ denotes a DP with concentration parameter, $\rho > 0$, and parametric base measure, G_0. The base measure G_0 is the distribution around which the DP is centered and the concentration parameter, ρ, reflects the variability around that base measure.

It is through the G_i's that the $f_i(r)$'s are clustered as can be seen from the stick breaking construction representation of the NDP: $Q = \sum_{k=1}^{\infty} \pi_k^* \delta_{G_k^*}(\cdot)$. In this representation, π_k^* represents the probability that a school is assigned to the k-th mixing measure, G_k^*, $\delta_\cdot(\cdot)$ is the Dirac delta function and $G_k^* = \sum_{l=1}^{\infty} w_{lk}^* \delta_{\boldsymbol{\theta}_{lk}^*}(\cdot)$ is itself composed of weights, w_{lk}^* and associated atoms $\boldsymbol{\theta}_{lk}^*$. This hierarchy of distributions, weights and atoms provides a framework that flexibly identifies clusters of schools, and also flexibly estimates the intensity function representing the spatial distribution of FFRs surrounding schools for each cluster.

Combined altogether, the hierarchical formulation of our model is:

$$
\begin{aligned}
\{r_{ij}; j = 1, ..., n_i\} &\overset{ind}{\sim} IPP(\lambda_i(r)), \quad i = 1, \ldots, N \\
\lambda_i(r) &= \gamma_i f_i(r) \\
f_i(r) &= \int \mathcal{K}(r|\boldsymbol{\theta}) dG_i(\boldsymbol{\theta}) \\
G_i &\overset{iid}{\sim} Q \\
Q &\sim DP(\alpha, DP(\rho, G_0)),
\end{aligned}
\tag{4.3}
$$

and, as previously noted, the γ_i are nuisance parameters that do not influence the estimation of the intensity functions, which are of primary interest.

In our analysis of FFR exposure around California public high schools we transform the school-FFR distances from $(0, R) \to \mathbb{R}$ using a probit function to create the modified distances $r'_{ij} = \Phi^{-1}(r_{ij}/R)$. We make this transformation in order to use a normal mixing kernel and corresponding normal-inverse-chi square base measure, $G_0 = N(0, \sigma) \times \text{Inv} - \chi^2(1, 1)$, in order to facilitate computation. Similarly, we place informative Gamma priors, on the concentration parameters, $\alpha, \rho \sim \text{Gamma}(10, 10)$, to encode our *a priori* belief that there should be a small number of clusters, in line with similar work (*Ishwaran and James*, 2001; *Gelman et al.*, 2013; *Rodriguez et al.*, 2008).

4.3.2 Health Outcomes Model

In the second stage of the analysis, we examine whether spatial distributions of FFRs around a school are associated with obesity of children in the school. For this, we use results from the clustering model (4.3) as input in to a regression model where child obesity is the outcome.

The simplest way to do this would to be to assign a cluster label to each school and use it as a covariate in the health outcome regression. However, proceeding

in this fashion would not account for the uncertainty in the cluster assignment and would not exploit all the information in the posterior samples that are generated while fitting the clustering model of Section 4.3.1. To address these issues, we propose two alternative approaches.

4.3.2.1 Consensus generalized linear model (CGLM)

The first approach *controls* uncertainty in the cluster labels by using in the health outcome model only the schools for which the cluster label is known with greater relative certainty, as follows. First, we derive cluster assignment labels for each school using the posterior samples and a loss function in a decision theoretic framework (*Wade et al.*, 2018). Specifically, we use the variation of information (VI) loss function to determine the optimal cluster configuration, which simultaneously identifies both the number of clusters and cluster labels for the schools. This approach finds the posterior sample that produces the minimal loss, and uses the number of clusters and cluster assignments in that sample to assign labels to schools– thus deriving, essentially, a "point estimate" for the discrete/categorical cluster assignment. We refer to this point estimate as the 'mode' cluster label.

Second, we identify schools with low uncertainty in the cluster label. In addition to the mode cluster configuration, the method also produces 95% uncertainty bounds for both the number of clusters and for the cluster labels for each school, yielding three additional cluster configurations (for a total of four including the mode). Compared to typical "upper" and "lower" bounds, the method provides three bounds according to the combination of loss metric value and the number of clusters, the latter of which can vary from one configuration to another. Since, as noted earlier, employing the cluster labels from the single point estimate as a predictor in the health outcome model would ignore the uncertainty associated with the cluster assignment label, our

goal here is to take into account all four cluster configurations to control or reduce the uncertainty. One possibility is to use each of the four assignments in separate health outcome regression models, and subsequently fuse together their results. However, fusing those four models may entail fusing models with potentially a different number of clusters. Thus, rather than using each of the four alternative cluster labels in the health outcome models, we restrict the health outcome analysis to the set of schools that are assigned to the same cluster across the four different cluster assignments. Note that this is only possible when clusters are well identified and posterior samples do not exhibit label-switching across iterations – as is our case – or a post-processing step that adjusts for label switching has been run (*Gelman et al.*, 2013; *Rodríguez and Walker*, 2014; *Stephens*, 2000; *Papastamoulis*, 2016). These conditions ensure that cluster labels are consistent across configurations and, consequently, taking the intersection has a consistent meaning.

In summary, the CGLM approach addresses the uncertainty in the cluster label assignment by taking the intersection of the four cluster labels to arrive at a more conservative (less uncertain) estimate of the schools' cluster assignment. This reduction of uncertainty in cluster assignment comes at the cost of sample size, as schools will be included or excluded from the health regression model according to whether they fall within said intersection or not. Despite this loss of sample size, the key advantages of this approach are that this enables a straightforward analysis, as the intersection of the four cluster configurations yields a single cluster assignment for each school that can be used as a categorical covariate in the health outcome model, and the cluster assignment is more precise than the single "point estimate" label in the entire sample.

To define the CGLM outcome model and enable us to distinguish the association between the quantity of FFRs and obesity from that of the FFRs' spatial distri-

bution, we bring back into consideration the schools that had zero FFRs within 1 mile. Let $C_{i,k} = I(\text{ith school belongs to cluster k})$, $k = 1, ..., K$, denote the cluster to which the i-th school is assigned ($K = 6$ in our analysis). In addition, define $Q_{i,m}, m = 0, ..., M$ a set of indicator variables that treat the number of FFRs as a categorical variable. In particular $Q_{i,0} = I(n_i = 0)$, with the other categories being $n_i = 2; n_i = 3; n_i = 4; n_i \in \{5, 6, 7\}$, and $n_i > 7$. This categorical representation is used given the distribution of FFRs and the lack of linearity in the association between FFR quantity and the odds of obesity. We note that the cluster indicators are only available for schools with $n_i > 0$, and that the schools included in this model are those with $n_i = 0$ or $n_i > 0$ that are determined by the consensus approach discussed above to have a cluster label with relatively higher certainty. This set of schools is denoted as $\mathbb{D}_{Consensus}$.

The CGLM model linking obesity in schoolchildren to exposure to FFRs and other school neighborhood characteristics is thus:

$$\text{logit}(\text{p}_{i'}) = \left\{ \sum_{m=0}^{M} Q_{i',m}\zeta_m \right\} + I(n_{i'} > 0) \left\{ \sum_{k=1}^{K} C_{i',k}\xi_k \right\} + \boldsymbol{Z}_{i'}^T\boldsymbol{\beta} \qquad i' \in \mathbb{D}_{consensus} \tag{4.4}$$

In (4.4), $\text{p}_{i'}$ denotes the proportion of obese 9th grade students at the i'th high school, ζ_m and ξ_k are quantity- and cluster-specific coefficients, respectively, and $\boldsymbol{Z}_{i'}$ is a vector of school characteristics without an intercept term. Specifically $\boldsymbol{Z}_{i'}$ includes the categorical variable indicating the racial majority of the children enrolled in the school; an indicator denoting whether school i' is a charter school; the school i''s census tract median household income centered at the overall state median income and scaled by 33,000; the proportion of adults who have $\geq$ 16 years of education within the school's census tract centered by the state census tract average; and the

urbanicity level near the school, classified as rural, suburban or urban. Given the parametrization of the covariates, the reference category when $Z_i = 0$ is a suburban high school, with a majority white student population, with the average percent of college educated adults and median census tract household income.

4.3.2.2 Bayesian Kernel Machine Regression (BMKR)

In the second approach, we move away from using a categorical predictor in the health outcome model, and instead use BKMR, as explained below, to input the probabilities that any pair of schools belong to the same cluster into the health outcome model. This has several advantages compared to approaches that use a single set of cluster labels for schools (e.g., the mode) in the heath outcome model. When a categorical predictor, denoting discrete groups of schools' FFR spatial profiles, is used in the health outcome model, we estimate the effect of each of these spatial profiles by borrowing information only from the schools within the same cluster. Hence, when we estimate health effects in this fashion, we are only borrowing information from a limited number of schools, those that belong to the same disjoint subset of schools, i.e., the clusters. Moreover, as described above in the CGLM, additional steps are needed to account for cluster assignment uncertainty. In contrast to the CGLM specifically, which controls uncertainty by restricting the sample, the BKMR could utilize all the schools in the NDP.

The BKMR approach proceeds as follows. First, to each school i, $i = 1, \ldots, N$, in the clustering sample, we associate the N-dimensional row-vector $\mathbf{P}_i$, i-th row of the co-clustering probability matrix $\mathbf{P}$. This matrix is constructed by averaging, for each i, the indicators $I(\text{school } i \text{ and } j \text{ are in the same cluster})$, for $j \neq i$ across the posterior samples, with the i-th element of $\mathbf{P}_i$ set equal to 1 by convention. By using the N-dimensional vector $\mathbf{P}_i$ as a predictor in the health outcome model, we allow all schools, including those that would not be assigned to the same cluster as

school i, to contribute to the estimation of the effect of the spatial distribution of FFRs surrounding the school on its schoolchildren's odds of obesity. In other words, in this approach we remove the hard boundaries that separate schools into disjoint subsets, and enable sharing of information about schools' FFR-spatial profiles across all schools, albeit the contribution of schools with more (compared to less) similar profiles is given higher weight.

Second, rather than linking the log-odds of obesity at school i to the N-dimensional row-vector $\mathbf{P}_i$, directly, in the logistic regression for obesity, we include as a *predictor* the scalar $h(\mathbf{P}_i)$ for school i. Since schools that have large probabilities of being co-clustered are likely to have similar N-dimensional row-vectors, while schools that are not very likely to be co-clustered are associated with more dissimilar N-dimensional row vectors, we model the N random terms $h(\mathbf{P}_i)$, $i = 1, \ldots, N$ jointly. Specifically, we model $(h(\mathbf{P}_1), \ldots, h(\mathbf{P}_N))$ as a finite-dimensional realization of a Gaussian process with mean $\mathbf{0}$ and Gaussian covariance function $\kappa(\cdot, \cdot; \phi, \sigma^2)$. The covariance function $\kappa(\cdot, \cdot; \phi, \sigma^2)$ is evaluated using as distance between any two N-dimensional row-vectors $\mathbf{P}_i$ and $\mathbf{P}_j$ the Euclidean distance, whereas the two parameters ϕ and σ^2 encode, respectively, the range of the correlation, e.g. the distance at which the correlation between any two N-dimensional random vectors is essentially negligible, and the marginal variance of each $h(\mathbf{P}_i)$. This approach borrows from the environmental epidemiological literature where researchers are often confronted with the issue of having multiple, potentially correlated, high-dimensional exposures.

As with the CGLM, we incorporate into the outcome model all observations with 0 FFRs within the mile, and use the indicators for quantity of FFRs nearby, $Q_{i,m}$, to distinguish the effect on obesity associated with the quantity of FFRs from their

spatial distribution. Altogether, the second health outcome model is:

$$\text{logit}(\text{p}_{i'}) = Q_{i',0}\zeta_0 + I(n_{i'} > 0)\left\{\tilde{\alpha} + h(\boldsymbol{P}_i) + \sum_{m=0}^{M} Q_{i',m}\zeta_m\right\} + \boldsymbol{Z}_{i'}^T\boldsymbol{\beta} \quad i' \in \mathbb{D}_{Full} \tag{4.5}$$

$$h(\cdot) \sim \mathcal{GP}(\mathbf{0}, \kappa(\cdot,\cdot|\phi,\sigma^2)).$$

In (4.5), $\mathbb{D}_{Full}$ is the index set for schools with zero FFRs in addition to the full set of N schools used in the first stage model. Additionally, $\tilde{\alpha}$ denotes the intercept for schools with at least one FFR, and $h(\boldsymbol{P}_{i'})$, $i' \in \mathbb{D}_{BKMR}$, indicates a school-specific random intercept. The other components of the model in (4.5) - $\boldsymbol{\beta}, \boldsymbol{Z}_{i'}$ - have the same definition and interpretation as before.

For comparative purposes, we fit the BKMR to both datasets, $\mathbb{D}_{Consensus}$ and $\mathbb{D}_{Full}$. Similarly, we also fit a logistic regression to schools in the $\mathbb{D}_{Full}$ set using the same parametrization as in (4.4). In this latter case, the mode cluster assignment was used to determine cluster specific indicators. This model is hereafter referred to as the Mode GLM (MGLM). In all models, $\boldsymbol{\beta}$, ζ_m,$m = 1, ..., M$ and ξ_k, $k = 1, ...K$ are given flat improper priors, while in the BKMR, ϕ and σ^2 are each given informative folded Normal$(1, 3)$ priors to accommodate known identifiability issues (*Zhang*, 2004). These informative priors were chosen after initial runs with uniform priors on larger intervals of $\mathbb{R}^+$ for both parameters showed that posterior samples were contained in the $(0, 1)$ interval.

4.3.3 Estimation

As both the clustering model and the health outcome models that we have proposed in Section 4.3.2 are specified within a Bayesian framework, inference on all model parameters are obtained through the posterior distribution, which we approximate using a Markov chain Monte Carlo (MCMC) algorithm. For our NDP cluster-

ing model we use the blocked Gibbs sampler as described in *Rodriguez et al.* (2008), truncating the summations for the inner and outer DPs using $L = 30$ and $K = 35$, respectively - a choice based on logic similar to that discussed in *Rodriguez et al.* (2008). This model fitting routine is implemented within our `bendr` (*Peterson*, 2020a) `R` package. The health outcome models are fit using the Hamiltonian Monte Carlo variant sampler implemented in `stan` (*Carpenter et al.*, 2016) via `rstan` (BKMR) (*Stan Development Team*, 2020) and `rstanarm` (CGLM) (*Stan Development Team*, 2016). All model fitting was performed within `R` (v3.6.1) (*R Core Team*, 2013) on a Linux Centos 7 operating system with 2x3.0 GHz Intel Xeon Gold 6154 processors.

For the NDP model, 250,000 samples were drawn from the posterior distribution, with 240,000 iterations discarded for burn-in and the last 10,000 iterations thinned by 3 to reduce auto-correlation for a total of 3,333 posterior samples used for inference. The length of the burn-in period and thinning were determined by inspecting trace plots for various model parameters and by computing Raftery's diagnostic statistic (*Raftery and Lewis*, 1995). For all health outcomes models, we ran 4 independent chains, using different initial values, each ran for 2000 iterations. For each chain, we kept 1000 samples after burn-in, for a total of 4000 posterior samples that we employed for posterior inference. Convergence was assessed via split $\hat{R}$ (*Gelman et al.*, 2013) and visual inspection of trace plots.

In fitting the proposed health outcome models, we exclude 17 schools previously included the clustering because they are missing outcome information. As discussed in Section 4.3.2, the outcome models include an additional 446 high schools with no FFRs within the 1 mile radius, which were not considered in the clustering model. Cluster labels that are included as predictors in the CGLM and MGLM were derived using the `mcclust.ext2` package in `R` (*Wade*, 2015). The same `R` package was employed to estimate the posterior assignment credible bounds with VI loss function as detailed in *Wade et al.* (2018). We take the intersection over the horizontal,

upper and lower bounds as described in Section 4.3.2 to arrive at our cluster assignment predictors for the CGLM, while the mode assignments are taken "as-is" for the MGLM.

For the NDP, posterior medians, inter-quartile ranges (IQRs) and 95% credible intervals are calculated for the intensity function parameters, $(\mu_{lk}, \sigma^2_{lk})^*$, π^*_k as well as the probability of co-cluster membership $\boldsymbol{P}$ as described in Section 4.3.2. The $f(r)$ were constructed over a fine grid of equally spaced values in $\mathbb{R}$ representing the distances of a BEF from a school, combining the K clusters and the L within-cluster components at each distance. Since, for computational purposes, we transformed the school distances to FFRs from the interval $(0, R)$ to $\mathbb{R}$, when inferring upon the actual densities $f(r), r \in (0, R)$, we back-transform them onto the $(0, R)$ domain using the inverse probit function, rescaling them by an empirically calculated proportionality constant.

In the health outcome models, posterior median and 95% credible intervals are calculated for regression coefficients $\boldsymbol{\beta}$, cluster effects $\xi_k, k = 1, ..., 6$, and $h(\boldsymbol{P}_i)$, $i = 1, \ldots, N$. Additionally, we calculate the effects of the quantity of FFRs as described in the following section.

4.3.3.1 Quantity Effect

Given that schools with $n_i = 0$ cannot be assigned a cluster for obvious reasons, the CGLM includes the non-standard interaction terms between quantity $n_{i'} > 0$ and cluster assignment. As in any model with interactions, describing the "main effect" of an exposure requires careful attention. In this case, the "main effect" of the quantity of FFRs depends on the cluster to which schools with $n_{i'} > 0$ were assigned. For each category indicator $Q_{i',m}$ with $m > 0$, we marginalize over the cluster assignment to define the probability of obesity given category $Q_{i',m}$; that is, the *quantity effect*,

holding $\boldsymbol{Z}_i = 0$ as defined in Section 4.3.2 , is

$$P(\text{Obesity} \mid Q_m, \text{Data}) = \sum_{k=1}^{K} w_k \text{inv-logit}(\zeta_m + \xi_k), \tag{4.6}$$

where w_k is the probability of a school being assigned to cluster k in the $\mathbb{D}_{CGLM}$ dataset. Note that for $n_i = 1$, there is no corresponding effect, ζ_1, as this is defined as the average cluster effect by the construction of the model in (4.4).

Similarly, for the BKMR we calculate the FFR quantity effect on schoolchild obesity, now averaging over the $h(\boldsymbol{P}_i)$'s, using the following probability expression:

$$P(\text{Obesity} \mid \zeta_m, \text{Data}) = \frac{1}{\mid \mathbb{D}_{BKMR} \mid} \sum_{i=1}^{|\mathbb{D}_{BKMR}|} \text{inv-logit}(\zeta_m + \widehat{h(\mathbf{P}_i)}), \tag{4.7}$$

where $\widehat{h(\mathbf{P}_i)}$ is the posterior mean of $h(\mathbf{P}_i)$.

4.4 Results

We now discuss results from both the clustering model, that aims to identify major patterns in the spatial distribution of FFRs in a 1 mile radius ($R = 1$) around schools, and the models that relate the spatial distribution of FFRs to the odds of obesity in schoolchildren.

4.4.1 Spatial Intensity Functions

The clustering model estimates six clusters with high probability, with the estimates of the cluster-assignment probabilities, π_k^*, beyond the first six effectively negligible when rounding to the hundredths place. The median density estimates, representing the likelihood of finding an FFR at a given distance from a school, are presented in Figure 4.2, along with the proportion of schools in each cluster. As the figure shows, clusters are labeled according to their mode's proximity to the school,

i.e. the cluster which estimates most FFRs to be located nearest to schools is labeled cluster 1, and so forth. A figure with the densities on the real line as estimated in the model along with 95% credible intervals is shown in Appendix A.

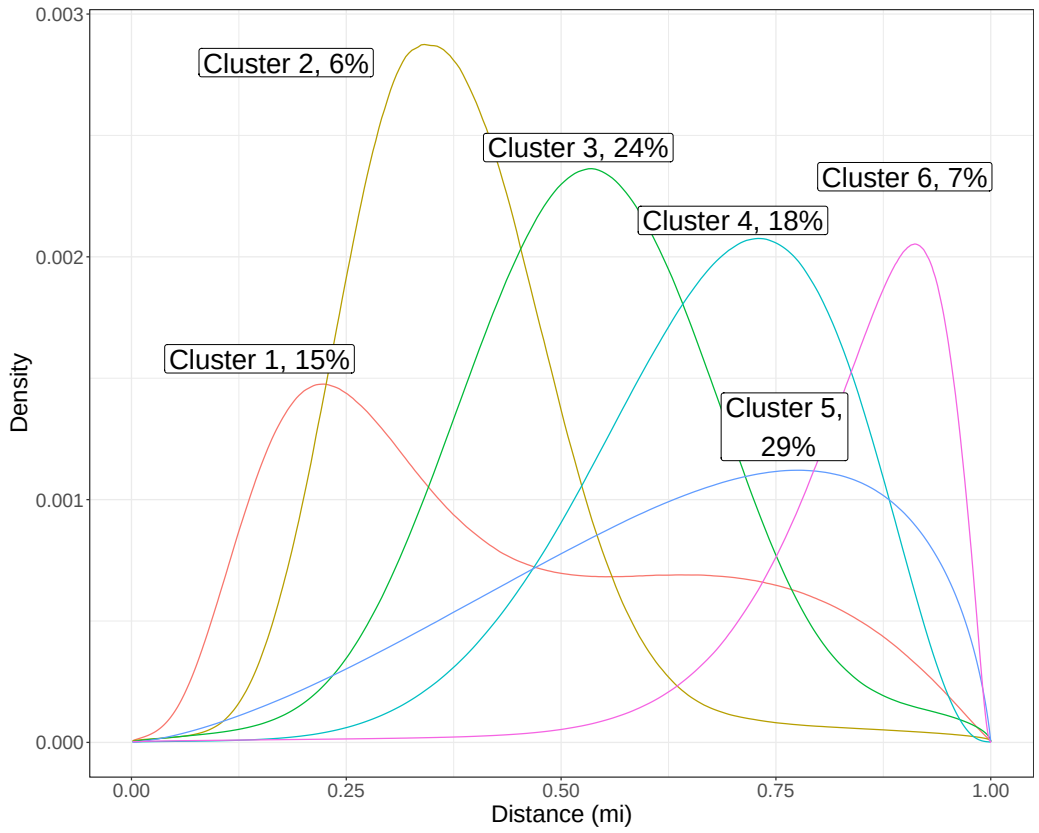


Figure 4.2: Estimate of cluster density functions $f_k^*(r)$, $k = 1, \ldots, 6$, with the estimated percent of schools within each cluster, π_k^*. The estimate here is taken to be the posterior median. The IQR for the percent of schools in each cluster are, for clusters 1 to 6, respectively: 3, 2, 4, 5, 5, and 2%

In order to visualize how distinctly the clustering model assigns schools into the six different clusters, Figure 4.3 presents a heat map of the co-clustering probability matrix $\mathbf{P}$. Note that in the Figure, for the sake of visualization, school indices are arranged so that schools with similar co-clustering probabilities are next to one another, as implemented in the algorithm described in *Rodriguez et al.* (2008)'s Supplementary Material. Examining the plot, we can clearly see the six clusters from left to right

followed by the remaining schools which the model cannot cluster consistently.

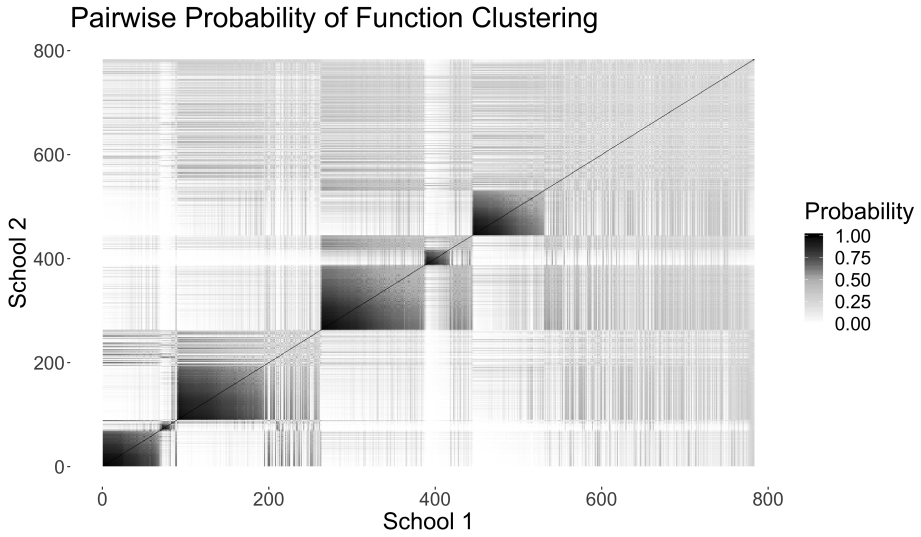


Figure 4.3: Heat map of co-clustering probabilities, that is, the probability that any two schools are assigned to the same cluster. The identity line may be interpreted as a school's probability of being clustered with itself. Although this probability is trivially equal to 1, for plotting purposes, in the figure this line is left equal to 0 to more clearly show the plot's line of symmetry.

Table A.3 presents summary statistics for the characteristics of the schools included in the six clusters identified, including a tabulation of the categorical variable describing the number of FFRs within a 1-mile of each school. In the table, we also include the characteristics of schools for high schools that have no FFRs within one mile of their location: we label this cluster as "Cluster 0". There is a weak association between a school's census tract median household income and cluster membership. While Cluster 1's median census tract median income is $55,200, Cluster 6's median census tract median income is $67,400. However, this patterning does not include Cluster 0, which has a lower median census tract median income of $53,900. A similar, though even weaker, pattern can be seen in the proportion of residents in the schools' census tracts with 16 or more years of education. Forty-four percent of schools in cluster 1 have majority of white students populations, whereas 38 have

predominately Latino students; for cluster 6, these percentages change to 58 and 16, respectively. Notably, all clusters contain schools across all urbanicity classification, and include schools with a varying number of FFRs. In other words, the mode cluster is not driven by FFR quantity or broader context (e.g., urbanicity) of the schools.

To assess whether the six identified clusters were geographically concentrated in one or more sub regions of California, and to investigate whether schools tended to co-cluster with nearby schools, we produced spatial plots of the co-clustering probabilities for a given school. Figure A.9 presents this plot for a school located in Southern California, identified in the map by a star symbol. As the figure shows, the schools that are more likely to be co-clustered with the selected school are not necessarily located nearby. Rather, as the first panel of the figure shows, most of the schools nearby the chosen school tend to have a probability smaller than 0.5 of being assigned to the same cluster.

4.4.2 Health Outcomes Models

In discussing the results of our proposed health outcomes models, we start by providing a description of the schools whose data are used. The proportion of students that are obese is similar in both the consensus and full datasets (Table 4.2), which is encouraging since schools are not excluded on the basis of the outcome in the consensus dataset. However, schools used to fit the consensus model are less likely to have few FFRs around them - only 21% have between 1 and 4 FFRs compared to 45% in the full dataset.

Turning to outcome model results, we'll discuss both second stage approaches on both the consensus and full datasets, starting with the consensus dataset. However, since the BKMR results mirror those of the CGLM, we'll focus on how these second-stage models reinforce one another rather than describing each individually.

As shown in Figure 4.4, we observe a monotonic decrease in the probability of

obesity as a function of the proximity of FFRs to the school, after adjusting for 1 mile radius quantity of FFRs. Specifically, according to the CGLM, children attending schools consistently assigned to cluster 6 have a 35% (95% CI: 33%,38%) probability of being overweight or obese, while, for other clusters, the lower bound estimate of the probability of obesity ranges from 37% to 40%. These results are consistent with the substantive expectation that students who are exposed to FFRs in the immediate environment around schools are more likely to be obese than they would be otherwise. As Figure 4.2 shows, the density of FFRs for schools in cluster 6 is greatest after 3 quarters of a mile, in explicit contrast to the other clusters which tend to have greater density of FFRs closer to the school. This finding supports prior work suggesting that zoning laws that restrict the placement of fast food restaurants could serve as possible population-level strategies to reduce child obesity (*Austin et al.*, 2005).

Figure 4.4 overlays the results of the CGLM and the BKMR models, demonstrating their agreement. The figure also shows that the BKMR provides additional information regarding the probability of obesity for children in each school. Beyond potential policy implications of the average obesity risk for children across the school food environment clusters, the school-level estimates can be used to prioritize individual schools for additional interventions.

Figure 4.5 shows the estimated probability of obesity as a function of the number of FFRs within a 1-mile radius of the schools, calculated as described in Section 4.3.3.1. As the figure indicates, there is a general agreement between the CGLM and BKMR models with respect to the negligible effect of the number of FFRs on obesity after adjusting for the FFRs' spatial distribution and other covariates. The only estimate that stands out from these analyses is the BKMR's estimate of lower obesity for children in schools with 5-7 FFRs nearby, as compared to zero FFRs - a counter intuitive result. However, it is possible that the greater number of FFRs implies greater variety of food choices, including healthier options. The data set in

this analysis does not contain information on the specific types of FFRs, beyond the number and location, thus not allowing us to examine this possibility.

As with the consensus data set, the results from fitting a GLM (using only the median cluster assignment) and the BKMR in the full data set are in agreement with each other, as shown in A. However, the results from the analysis on the full data set instead identify Cluster 2 as having the lowest probability of obesity, at 37% (95% CI: 36%,38%), with the probability for all other clusters near or above 40%. Differences in the association between the spatial distribution of FFRs near schools on child obesity, comparing the full and consensus data sets are likely due to the fact that the full dataset contains schools with more uncertain cluster assignments, and thus potentially more prone to miss-classification errors and thus bias in the associations. The quantity effects are similar in the consensus and full data set, and again agree between methods (see Figure A.13).

Finally, we compare and validate our models. Comparison to traditional competitor models by WAIC are shown in Table A.4, while posterior predictive checks are available Figure A.14. Both the BKMR and CGLM perform better by this metric than their more traditional counterparts regardless of what dataset they have been fit to.

	In Consensus	Not in Consensus	All
Proportion Obese	40.9 (33.3-47.4)	41.3 (34.1-48.2)	41.3 (33.9-48)
FFR Quantity within 1 mile			
[1,4]	21	55	45
$\geq$5	79	45	55
Urbanicity			
Rural	8	11	10
Sub-Urban	35	47	44
Urban	57	42	46
Majority Race			
African American	1	1	1
Asian	5	4	4
Hispanic	30	28	29
No Majority	16	13	14
White	49	53	52
Income[1] (1,000 USD)	60.4 (43.2-78.4)	61.4 (46.2-82.9)	61.2 (45.3-81.5)
Education[2]	25.3 (24.3-26.7)	25.4 (24.2-27)	25.4 (24.2-26.9)

Table 4.2: Descriptive Statistics for Schools Analyzed using the Consensus GLM vs. not. FFR = Fast Food Restaurant. All numerical values for categorical rows are the column percentage within the left row heading. [1] Median income of the households within the school's census tract. [2] proportion of individuals with $\geq$ 16 years of education within the school's census tract.

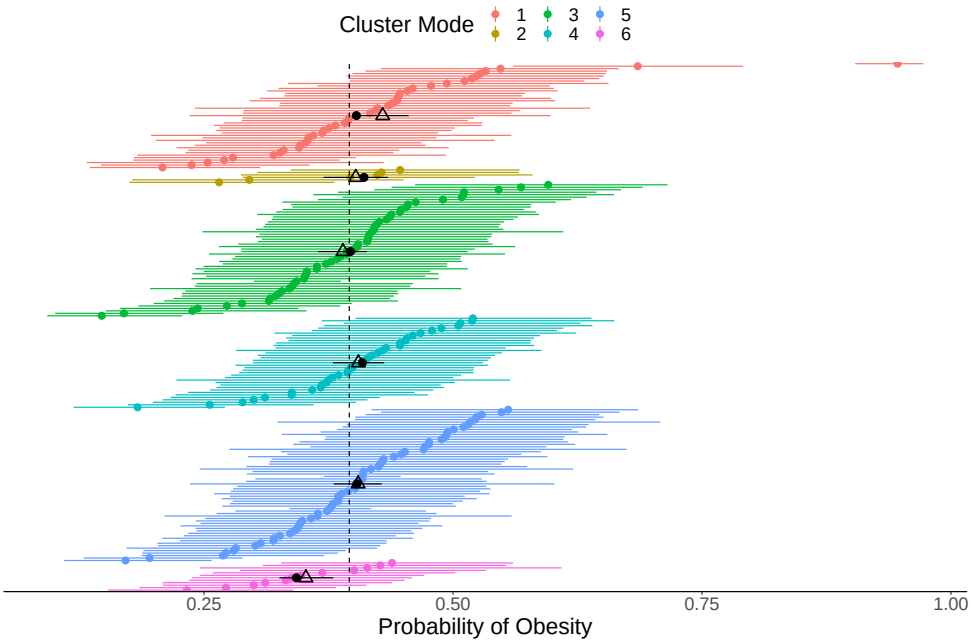


Figure 4.4: Probability of obesity in relation to fast food restaurant (FFR) proximity. Estimates from the Bayesian Kernel Machine Regression (BKMR) are shown for each school (dot), along with 95 % credible intervals (line), and are colored according to the cluster mode assignment. Dark dot represents the overall median probability of obesity for children attending schools in the given cluster. Triangles (and horizontal black line) denote the median posterior probability of obesity for children attending schools in each cluster estimated from the consensus GLM (CGLM) along with the 95% credible interval interval. The reference dotted vertical line is the posterior mean probability of obesity at a majority White suburban high school with at least one FFR within a mile of the school's location. BKMR and CGLM results are estimated using the consensus data set.

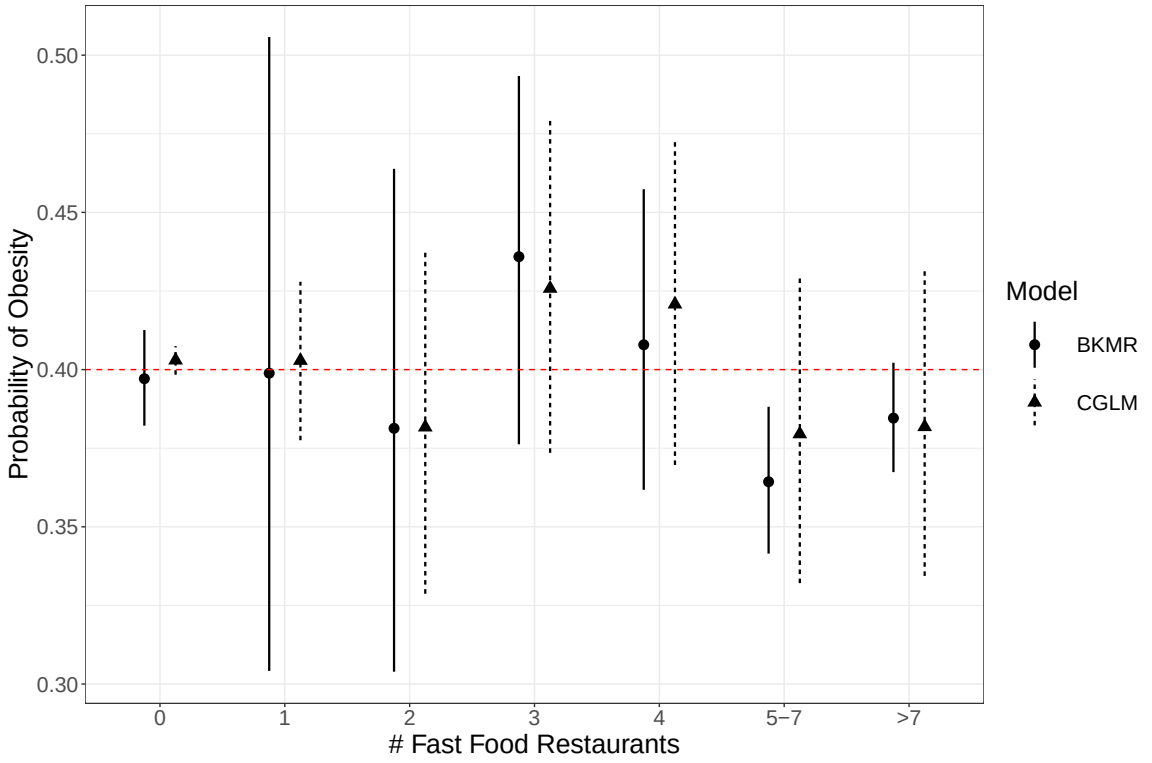


Figure 4.5: Posterior probability of obesity and 95% credible intervals according to the number of FFRs surrounding a school, adjusted for the effect of proximity of FFRs. Results refer to analysis performed on the consensus dataset, as estimated by the two models.

4.5 Discussion

In this work we have presented a two-stage modeling strategy that aims to provide epidemiological and social science investigators with a tool that permits them to both identify links between exposure to specific features of the built environment and health outcomes as well as identify those subjects at greatest risk of negative health outcomes. In the first stage, our goal is to identify major patterns in the spatial distribution of FFRs around a school and group schools together based on their surrounding food environment. The second stage links the spatial distribution of FFRs around schools to the likelihood that children in the schools are obese. This

work can be easily adapted to answer questions involving the association between other point-referenced amenities in the built environment and health outcomes, for example, availability of parks and measures of physical activity (*Evenson et al.*, 2016), depression (*Bojorquez and Ojeda-Revah*, 2018), or availability of social engagement destinations and cognition, among others (*Besser et al.*, 2018).

Our work breaks with previous approaches to quantify exposures to the built environment in several ways. First, we use a point pattern approach to model exposure to BEFs, which is not typically done in this type of literature. Second, our clustering model differs from other parametric clustering models used in built environment research in that no constraints are imposed on the number of clusters. Third, in addition to selecting the number of clusters in a data-adaptive fashion, our model estimates the cluster specific densities non-parametrically, thanks to the use of the NDP prior on the density function representing the distances between a given school and nearby FFRs. This modeling strategy allowed us to identify clusters of high schools in California that have a high number of FFRs in their immediate environment relative to their peers, and those that have FFRs farther way.

Our work also proposes approaches to incorporate output from a clustering method into a second stage regression model; namely, by using either a decision-theoretic framework to control uncertainty in cluster assignment, or using the posterior co-clustering probability matrix to borrow information across observations without needing to reduce the exposure information to discrete exposure groups. The latter approach extends the uses of kernel machine regression to applications in the built environment, beyond those used to examine health effects of chemical contaminants (*Bobb et al.*, 2015). Through our proposed second-stage health outcome models, we incorporated information on the spatial distribution of point-pattern amenities into a model for child obesity. These approaches identified that, independent of the quantity of FFRs within a mile of California high schools, children attending schools with

FFRs further away had lower obesity risk. Though the differences in obesity risk between food environment clusters were relatively small, it is well understood that ubiquitous exposures, even if they have small individual-level effects, can result in large population health impacts (*Rose*, 2001).

The second stage models have different advantages, disadvantages and ways of incorporating BEF exposure information obtained from the first stage. Specifically, the CGLM has the potential to suffer from selection bias by using only the subjects with highest certainty in their class assignment. In our analysis, although the schools with higher uncertainty tended to have fewer outlets nearby, the excluded schools did not differ in terms of the outcome, thus minimizing selection bias concern upon conditioning by the number of FFRs in the second stage. Users of the CGLM should keep in mind this potential for selection based on the outcome, in which case inverse probability of selection into the second stage could be used. The BKMR can use all subjects, handles cluster assignment uncertainty by using the co-clustering probability matrix, and can provide more granular information about health outcome risk for each subject in a school through the posterior estimates of $h(\mathbf{P}_i)$. However, visualizing/interpreting the BKMR's rich set of output could be challenging. In our case, our goal was to compare the results of the analyses between the two methods and thus we used the mode cluster label to visualize the BKMR results. Other visualizations of the results may include displays of plots of the $\widehat{h(\boldsymbol{P}_i)}$ as a function of the L^2 norm of the co-clustering probabilities, $\boldsymbol{P}_i$ and $\boldsymbol{P}_j$ for a reference school j.

Pursuing methods that more comprehensively estimate or propagate the uncertainty associated with cluster assignment from a DP clustering approach through the health outcome analysis may be desirable, as neither the BKMR or CGLM fully do so. One possible solution would be to develop a method that allows for joint estimation of both the cluster specific densities as well as the cluster-associated outcome risk. Our current method is unable to easily embrace such a joint modeling approach

due to to both label switching and the varying number of assigned clusters across the MCMC iterations, yielding identifiability problems for the health outcomes models (*Gelman et al.*, 2013). This makes the goal of joint estimation more difficult and a promising subject of future work. One possible solution would be to incorporate the health outcome at the level at which the cluster is constructed. Adapting the Logistic Stick Breaking Process (*Ren et al.*, 2011) for example, could facilitate this goal. Nevertheless, we emphasize that, while the two-stage approach proposed here does not propagate uncertainty in cluster assignment in a standard fashion, this strategy still offers a number of benefits. For example, defining the exposure clusters independent of the outcome ensures a greater level of interpretability and conforms to substantive understanding of such clusters (*Nylund-Gibson et al.*, 2019). Furthermore, it offers a greater level of applicability for the clusters, as estimating them separately from the outcome means they can be used for more than one health outcome analysis.

Finally, we acknowledge two recent theoretical results pertaining to the NDP and DP, respectively. In the first, *Camerlenghi et al.* (2019) showed that the NDP estimates of two or more distributions can collapse or degenerate to a single estimated distribution in the case when there are ties in the observations across subjects (e.g. two schools that have the exact same school-FFR distance) or when the true underlying distributions for different clusters share atoms. In our case, this can lead to collapsing or merging of intensity function estimates and thus potentially a lower number of school clusters identified. In the case of ties, we believe this should not be of substantial concern in the application that we are proposing our methodology for, as ties rarely occur in the calculation of distances at sufficient precision (none occurred in our motivating dataset) and, if ties do occur, a small amount of normal random error can be added to the distances to quickly resolve this issue without significantly affecting inference. The question of whether a latent mixture component is shared between two normalized intensity functions is more difficult to answer. How-

ever, we believe that the variability in these kinds of data, as illustrated by Figure 4.1, provides evidence that this may be of less concern here. If a dataset exhibits less variability than the one examined here, then greater caution may be warranted.

The second theoretical result of note showed that the DP cannot consistently estimate the number of clusters when the concentration parameter is fixed (*Miller and Harrison*, 2013, 2014). The same authors suppose a similar result holds in the more general case of a random concentration parameter, though there is as yet no proof (*Miller and Harrison*, 2018). The challenges this unknown feature presents can be accommodated by the two different outcome models we presented. If one is unconcerned about the potential lack of consistency and willing to rely on the concentration parameter to correctly inform the appropriate number of clusters, then the CGLM will offer a standard interpretation that relies on the number of clusters being a consistent estimate of the truth. In contrast, should there be concern that the NDP cannot consistently estimate the correct number of clusters then the BKMR offers a better approach – as it does not rely on the concept of their being some definite number of clusters, but rather uses the matrix of co-clustering probabilities to provide information about differing levels of exposure.

Further extensions of the work presented should incorporate the spatial distribution of more than one BEF amenity. The built environment consists of many amenities beyond FFRs that could be co-located with FFRs, or have different spatial distributions leading to different 'mixtures' of amenities. Extending our model to higher dimensions could allow investigators to characterize joint exposure to multiple amenities and identify their corresponding relationship with health outcomes. In addition, model extensions that incorporate the spatial proximity of subjects, in our case schools, when forming clusters would be of interest. In our analyses, we selected schools that were far apart from one another to satisfy independence assumptions. The work presented here represents a first step in building relevant descriptions of

the built environment in which humans live and informing decisions as to how new built environments may be constructed in the future.

CHAPTER V

Bentobox: The Built Environment Network Objects Toolbox R package

5.1 Introduction: What is the `bentobox`?

The `bentobox` (Built Environment Network Objects Tool Box) package is a "package of packages" which cumulatively offer a collection of R data structures, modeling functions and accompanying visualization tools that facilitate the analysis of built environment data. Inspired by the `tidyverse` (*Wickham et al.*, 2019), which offers tools for more general data manipulation, and visualization, as well as `tidygraph` (*Pedersen*, 2018) which provides a straightforward user interface for working with relational data structures, the `bentobox` package builds on these two approaches to provide users a familiar API for those interested in working with and analyzing built environment data. The latter focus on analysis bears special mention, as the majority of the core bentobox packages are specifically focused on providing statistical methods for understanding built environment effects.

Analyzing the built environment, with respect to its impact on its inhabitants, requires a non-standard data structure that the `bentobox` is explicitly designed to address. Specifically, the standard "tidy" dataset of unique subject-measurement rows is augmented with additional tidy datasets of many distances, times, and other at-

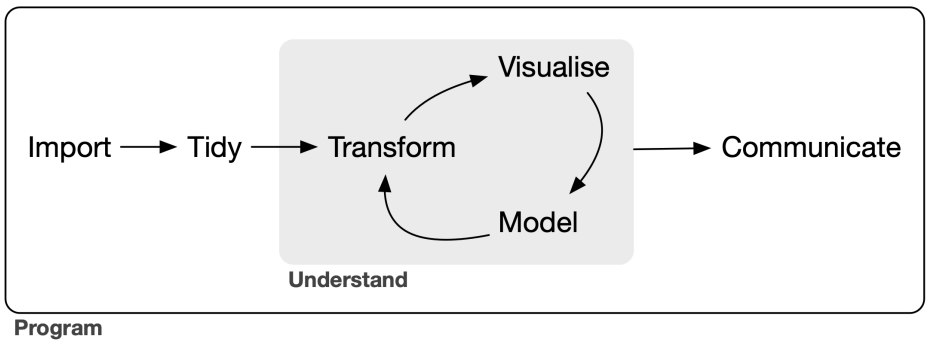


Figure 5.1: Statistical Workflow. Image credit to *Wickham and Grolemund* (2016).

tributes associated with those built environment features (BEFs), parks, restaurants or other amenities, in the environment and the subject at the time of measurement (See Figure 5.2 for an example). That is, the augmented datasets consist of unique subject-measurement BEF attributes. These attributes almost always include geographic and/or temporal components, as they are thought to be the first dimensions among which BEF effects may be observed *Baek et al.* (2016a).

In manipulating and modeling this non-standard data structure, the `bentobox` works seamlessly alongside other packages that are useful in this area, such as `osmdata` *Padgham et al.* (2017), for querying the location and shapefiles of BEFs as recorded in openstreetmap or `tidycensus`(*Walker et al.*, 2018) for subject outcome and covariate data at areal levels. In this paper I provide a brief description of where the `bentobox` fits into a built environment analysis framework and one example highlighting how data access via `osmdata` and data structures and models via the `bentobox` combine to provide investigators, policy makers, statisticians, and others interested in studying the built environment, an easy way to get started analyzing built environment data.

5.2 Context: Where does `bentobox` fit into an analysis?

As formulated in *Wickham and Grolemund* (2016), the typical statistical or "data-science" workflow consists of the following 6 high level steps as listed in Figure 5.1:

(1) Import (2) Tidy (3) Transform (4) Visualize (5) Model and (6) Communicate. While the `tidyverse` addresses these problems generally, providing functions that allow a user to read in an excel or csv file for example (Import), filter the subsequent dataset (transform) and then plot it (visualize), the `bentobox` builds on top of and in addition to the `tidyverse` in order to address similar and additional issues for users who are specifically analyzing built environment data. For example, while a user could still import built environment data using a `tidyverse` function or perhaps a function from `osmdata` or `tidycensus`, they could then join, *transform* and *visualize* the subject and built environment data using `rbenvo` *Peterson* (2020b), a `bentobox` package of built environment data structures. They could also then *model* the relationship between subjects and nearby BEFs using `rsstap`, `bendr` or `rstapDP`, each of which offer a different modeling approach for analyzing built environment data .

To briefly elaborate on each of these models, rsstap provides a penalized-spline approach towards estimating the effect of BEF exposure on health outcomes that will be demonstrated in section 3. `rstapDP` estimates a similar model as `rsstap` but allows for heterogeneity in the identified effects of the BEF, by placing a Dirichlet process prior *Gelman et al.* (2013) on the smoothing function's coefficients. Finally, `bendr` identifies clusters in the distributions of spatial exposure to a BEF of interest using the Nested Dirichlet Process *Rodriguez et al.* (2008) to avoid imposing parametric constraints on either the number of possible clusters, or the form of the BEF occurrence distribution. For further information see the documentation for each of these packages linked through the `bentobox` website.

Each of the packages offers a similar API for users, with `rbenvo` adopting many of the verbs from `tidyverse`'s `dplyr` package, and the modeling packages `rsstap`, `bendr` and `rstapDP`, each providing a familiar augmented formula for model fitting and adapted methods from the popular `lme4` or `stats` packages for obtaining model output. We'll demonstrate some of these in the following example.

5.3 Example: What does `bentobox` do?

In the following example, we'll use student obesity data obtained from the California Department of Education's Fitnessgram project and Fast Food Restaurant (FFR) location data from `osmdata` to show how the `bentobox` packages facilitate an analysis of built environment data. The full code for data import and preparation can be found here.

5.3.1 Import and Tidy

While the *Import* step is still technically carried out by packages in the `tidyverse` and `sf` packages, the `rbenvo::add_BEF` function does create a "tidy" dataset, by constructing the pairwise subject-FFR distance dataframe, which in the code below can be "activated" for display and further manipulation via the `rbenvo::activate` function (inspired by `tidygraph`).

```
bdf <- benvo(LA_schools, by='cdscode') %>%
  add_BEF(FFR, bef_id = 'osm_id') %>%
  activate(FFR)
```

5.3.2 Transform and Visualize

The "benvo" (Built Environment Object) returned by the above function can then be further manipulated and *transformed* via a subset of the familiar `dplyr` verbs including `mutate` and `filter`, as in the code snippet below, where we convert the calculated meters to kilometers and subset to include only those FFRs within 10km of a school.

```
bdf <- bdf %>%
  mutate(Distance = as.numeric(Distance/1E3)) %>%
```

```
filter(Distance<=10)
bdf
```

```
    Active df:  FFR
# A tibble: 489,034 x 3
   cdscode        osm_id   Distance
   <chr>          <chr>       <dbl>
 1 19101991933399 60713740     5.91
 2 19647330100289 60713740     9.89
 3 19647330100867 60713740     9.55
 4 19647330101683 60713740     5.48
 5 19647330102426 60713740     6.37
 6 19647330106435 60713740     9.36
 7 19647330108886 60713740     7.89
 8 19647330109439 60713740     3.22
 9 19647330110304 60713740     8.82
10 19647330111658 60713740     5.91
# ... with 489,024 more rows
```

Figure 5.2: Example subject-BEF augmented dataset

In order to *visualize* the spatial data, we can use "rbenvo" plot functions which serve as wrappers around the more sophisticated "ggmap" and "ggplot" functions for working with spatial data.

```
plot(bdf, 'map') + ggplot2::theme_bw() + ggplot2::theme_void()
```

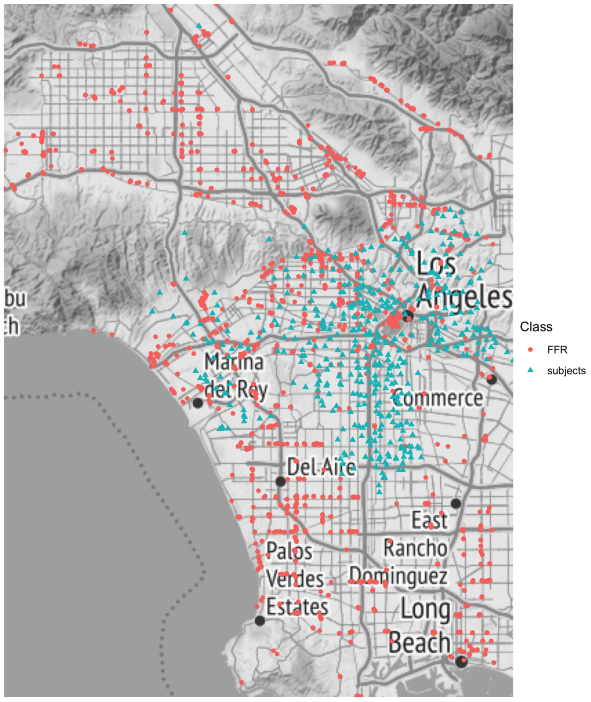


Figure 5.3: Sample map visualization via `rbenvo`

5.3.3 Model

Finally, to complete the general offerings of `bentobox`, we can fit or *model* one of the three models as listed in Section 5.2. To keep things simple, we'll fit a spatial temporal aggregated predictor model from `rsstap` *Peterson* (2020c), estimating the odds of student obesity at each school as a function of school or class-level covariates

and exposure to the nearby FFRs:

$$\begin{aligned}
\text{logit}(P(obesity_{ij}|b_i)) &= \alpha + I(Grade_{ij}=7)\delta_1 + I(Grade_{ij}=9)\delta_2 \\
&\quad + I(isCharterSchool_i)\delta_3 + f(FFR_i) + b_i \\
f(FFR_i) &= \sum_{d\in\mathcal{D}}\sum_{l=1}^{10}\beta_l\phi_l(d) \\
b_i &\sim N(0,\sigma_b^2) \\
\text{school } i &= 1,...,499 \\
\text{Grade} &= 1,...,n_i
\end{aligned} \tag{5.1}$$

```
fit <- sstap_glmer(cbind(NoStud_Obese, NoStud_NotObese) ~
                     Charter + Grade + sap(FFR) + (1|cdscode),
                     # sap = Spatial Aggregated Predictor
                     benvo=bdf,
                     family=binomial())
```

This is a naive model, intended more for demonstration purposes, than to offer anything authoritative, but nonetheless the estimate of the FFR exposure effect, seen in Figure 5.4, fits with what one might expect based on substantive reasoning alone: After adjusting for grade level and school charter status, the closer a FFR is to a school, the more likely the students at that school are expected to be obese.

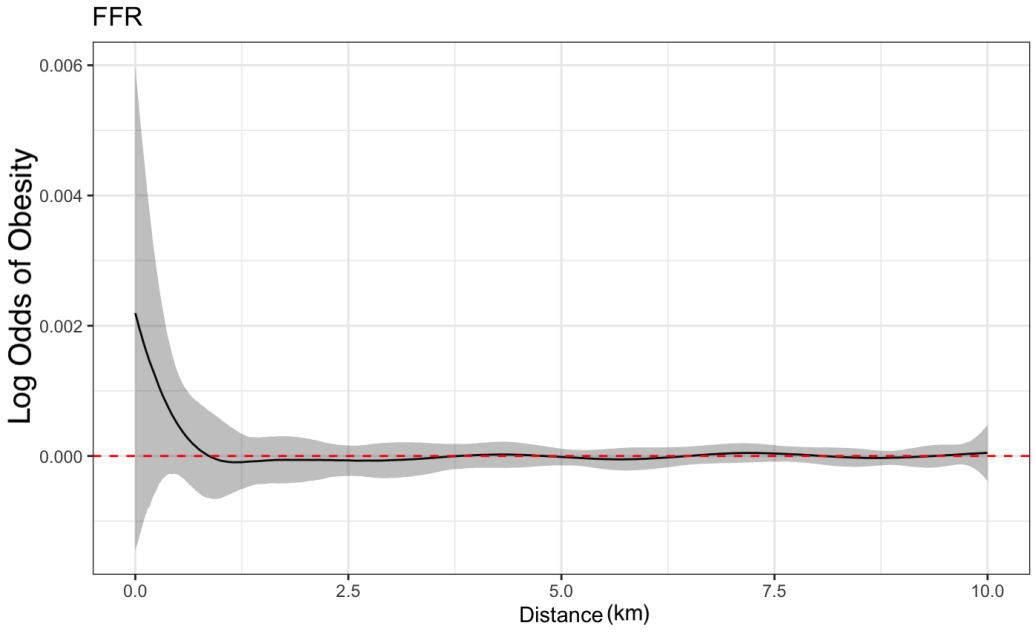


Figure 5.4: Sample FFR Efffect visualization via `rsstap`. Line is median estimate and ribbon represents the 95% pointwise credible interval.

5.4 Discussion

In closing, the `bentobox` package offers users a variety of tools and techniques to work with built environment data, offering a familiar API for the fluent `R` user. While there is undoubtedly still much that can be added to the `bentobox` in order to accommodate ever larger datasets and faster processing. As it currently stands, the `bentobox` ecosystem of software offers a strong and stable first step to giving users a variety of specialized tools for working with built environment data.

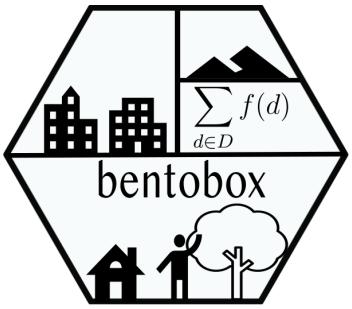


Figure 5.5: `bentobox` R hex

CHAPTER VI

Discussion

The availability of built environment data has increased dramatically over the past few decades. What was once unimaginable – knowing the locations of all Fast Food Restaurants across the state of California – is now easily and even publicly accessible via sources like OpenStreetMap or GoogleMaps ©. This dissertation sought to bring the questions these data naturally provoke, "how does where I live matter?", into sharper relief by constructing models that allow for the estimation of BEF effects on health outcomes. Whether estimating this effect across space and time (Chapters II,III) and/or identifying subjects at higher risk from exposure (Chapters III,IV), the methods developed in this dissertation provide investigators with new statistical and computational (Chapter V) tools with which to research this space. In this section I provide a final overview of the contributions from each of the cited chapters and present ideas for future work.

In Chapter II we took inspiration from *Heaton and Gelfand* (2011) to construct a spatio-temporal predictor that can be incorporated in standard regression frameworks allowing investigators to estimate the spatiotemporal functions and scales at which BEFs impact human health. Each of these represent their own novel methodological contribution to the literature. We examine in simulations the variability in estimates that result from using different choices of spatial exposure functions when the model

is both correctly and incorrectly specified. Our application to the North Carolina participants of the MESA cohort identified increased exposure to HFS resulted in a lower BMI on average when comparing between individuals. A main limitation of this work is the computational complexity required to evaluate the parametric nonlinear function that reflects the decay or accumulation of BEF effect across space and time. We demonstrated one solution to this issue in Chapter III which replaces the parametric nonlinear function with a linear combination of basis functions. By exploiting the linearity in this formulation, we are able to avoid the multiple aggregations across subjects' varying distance exposure during estimation, instead aggregating all distances before model-fitting, resulting in huge gains in sampling speed. However, the non-parametric approach loses the intuitive functional constraints enjoyed by the former. Should this be of primary importance, future work could look to augment the non-parametric approach by enforcing monotonicity on the basis function expansion coefficients.

The non-parametric approach offers further opportunities for future work. One such path could involve loosening the assumption of additivity present in the STAP models' construction of spatiotemporal exposure. While this assumption facilitates ease of computation and interpretation of the model, it is unrealistic and represents a limitation in the current construction: the first FFR at distance d likely has a much larger effect than the 50th FFR at the same (or even nearby) distance d.

Other directions for this family of models could include the construction of an appropriately structured group penalty since, as the number of spatio-temporal BEF predictors grow, the number of regression coefficients involved will likewise grow very quickly. Incorporating a realistic sense of the sparsity of BEF effects will enable new insights in the relative importance of one BEF type relative to its' counterparts.

Chapter III itself presents one extension to this family of models non-parametric STAP models, and it is to the STAP-DP model to which we now turn our attention.

By placing a Dirichlet Process Mixture prior on the regression coefficients involved in the basis function expansion we showed how one could flexibly estimate multiple curves, each belonging to a different segment of the population. Our main contribution in methodology here lies in the identification of these sub-populations with differing BEF effects, as well as their corresponding effects themselves. We study this model's performance in simulation, showing that both the distribution of distances as well as the difference in effect size between clusters can impact inference. Finally, our application identifies the presence of two groups' differing risk from FFR exposure in a sample of Los Angeles schools' census data from 2001-2008. Future work in this area includes increasing the dimensionality of the BEF exposure to include a temporal component, as well as spatial through a tensor product of the spline basis function expansion. Alternatively, or additionally, this model could be extended for other exponential family outcome data where the posterior of the spline regression coefficients are not available in closed form.

In Chapter IV, our contribution consisted of both demonstrating how the Nested Dirichlet Process could be adapted to flexibly estimate an unknown number of clustered BEF spatial distributions as well as how to conduct a second-stage analysis of the NDP output in order to infer cluster effects in a coherent manner. Our application to CA high schools found 6 clusters in FFR exposure across 1 mile, and a significantly decreased risk of obesity amongst those consistently clustered amongst those individuals attending schools with the most distant FFRs. Future work in this area could include the incorporation of multiple BEFs for multiple dimensions of density estimation. Alternatively, adapting our approach from a two-stage analysis to a joint analysis could allow for full propagation of uncertainty in cluster assignment to risk association estimates.

In Chapter V we demonstrated a suite of software packages nested within a master package titled `bentobox`, that implement the various methods discussed in this

dissertation as well as a host of auxiliary tools that facilitate the manipulation, organization and visualization of BEF data. We demonstrated how one could use publicly available data via the California Department of Education and OpenStreet Map ©to quickly and easily estimate the impact of BEFs on human health. Areas for future development of the `bentobox` include incorporating better routines for handling massive and on disk-only spatio-temporal datasets.

With the mass of built environment data only expected to grow as the economy becomes increasingly digitized, methodological challenges unique to this setting will grow in proportion. Handling issues with high dimensionality, spatio-temporal dependence structures and the identification of heterogeneity are all trends that analysts in this field have to confront on a day-to-day basis. This dissertation presents modeling frameworks that address these practical challenges through the course of 3 methodological projects and their corresponding software implementations. What's more, they lay the foundation for future work in this space, so that human understanding of how the places where we live may impact us can continue to grow.

APPENDICES

APPENDIX A

Appendix

A.1 Chapter 2 Supplementary Material

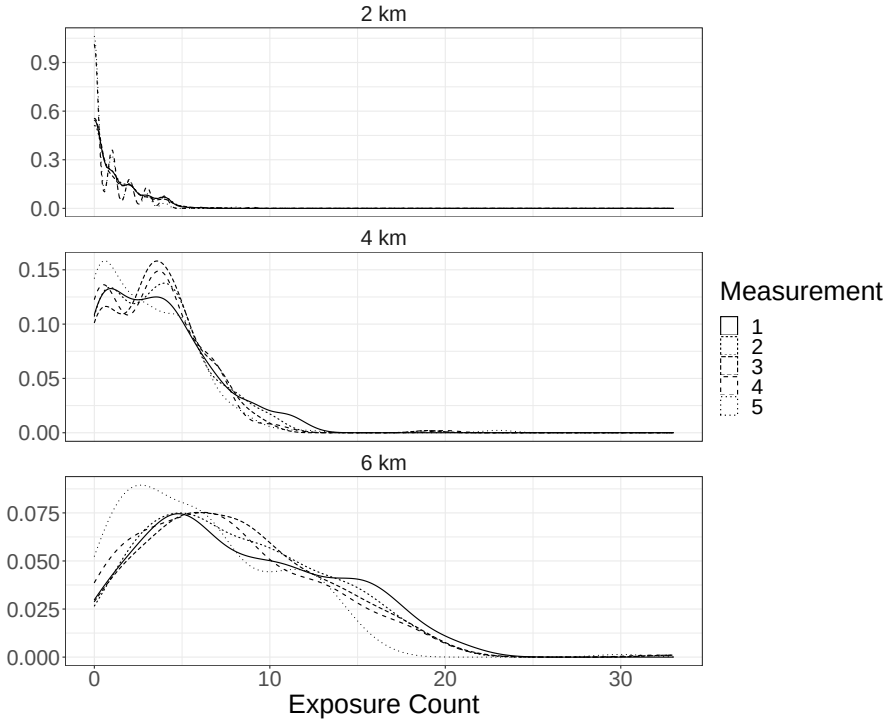


Figure A.1: Density Estimate of Healthy Food Store Exposure for North Carolina MESA participants across measurements- indicated by line type- and paneled by buffer sizes. The panel title indicates the buffer size.

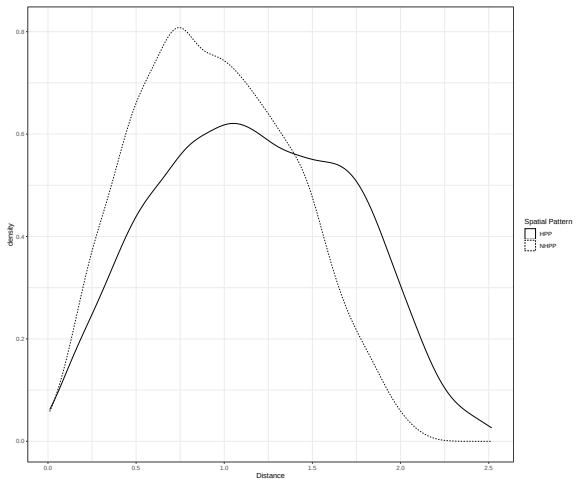
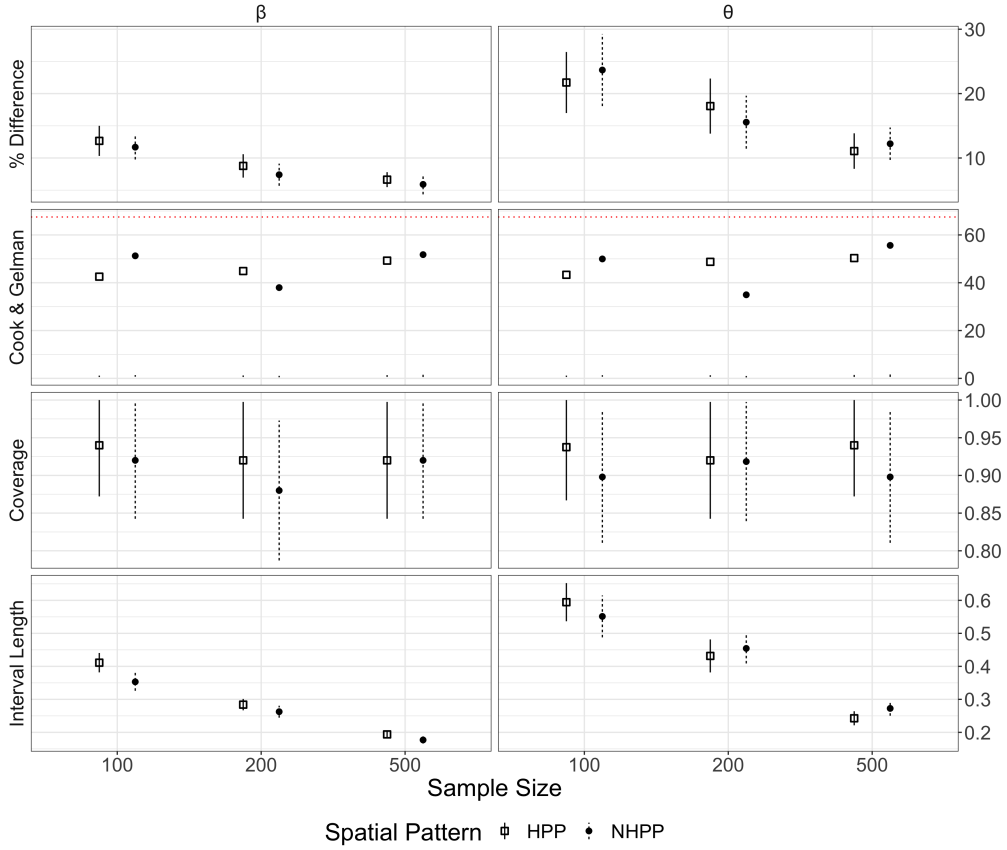


Figure A.2: Simulated Spatial Pattern Distance Distributions

Characteristic	**N = 311**[1]
Sex	
Female	170 (55%)
Male	141 (45%)
Race	
Black	129 (41%)
Hispanic	2 (0.6%)
White	180 (58%)
# Of Follow Up	
1	2 (0.6%)
2	16 (5.1%)
3	59 (19%)
4	234 (75%)
Education	
High School/GED (or less)	75 (24%)
Some College	102 (33%)
Bachelors	134 (43%)
Income (1000s USD)	
<12	17 (5.5%)
12-24	56 (18%)
25-39	51 (16%)
40-74	101 (32%)
>75	86 (28%)
Married	195 (63%)
History of Cancer	34 (11%)
Smoking Status	
Never	131 (42%)
Former	127 (41%)
Current	53 (17%)
Age at Exam	60 (53, 68)
Body Mass Index (kg/m^2)	28.0 (25.0, 31.6)

Table A.1: MESA Subjects descriptive statistics at baseline. [1]Statistics presented: n (%) ; median (IQR).

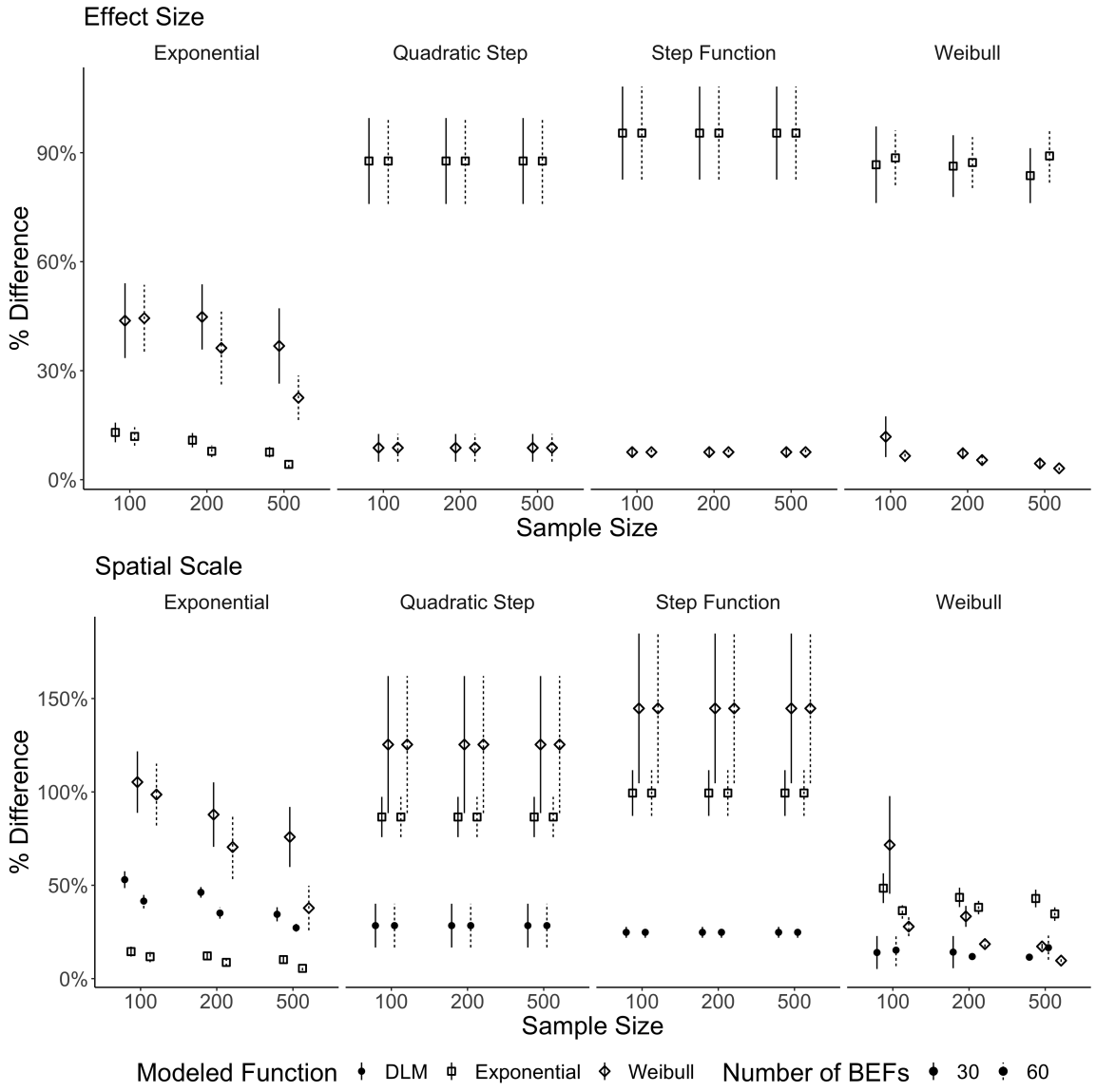
Figure A.3: Conservative prior simulation results Evaluated by (Top Row) Absolute Difference, (2nd Row) Calibration Statistic (3rd Row) Coverage and (4th) Interval Length across Sample Size and Spatial Pattern. For all plots but Cook & Gelman, dots and intervals indicate median, 95% credible interval respectively.



Reproducible code for the simulations can be found at:

https://github.com/apeterson91/STAPSimulations.

Figure A.4: Percent difference in median estimate of (Top) effect size β and (Bottom) spatial scale $d^\star$ from simulations varying information in sample size, generated spatial exposure function and modeled spatial exposure function using conservative prior.Panel title indicates the generating spatial exposure function, while dot shape indicates the median absolute difference for the modeled spatial exposure function. Line width is the 95% credible interval.



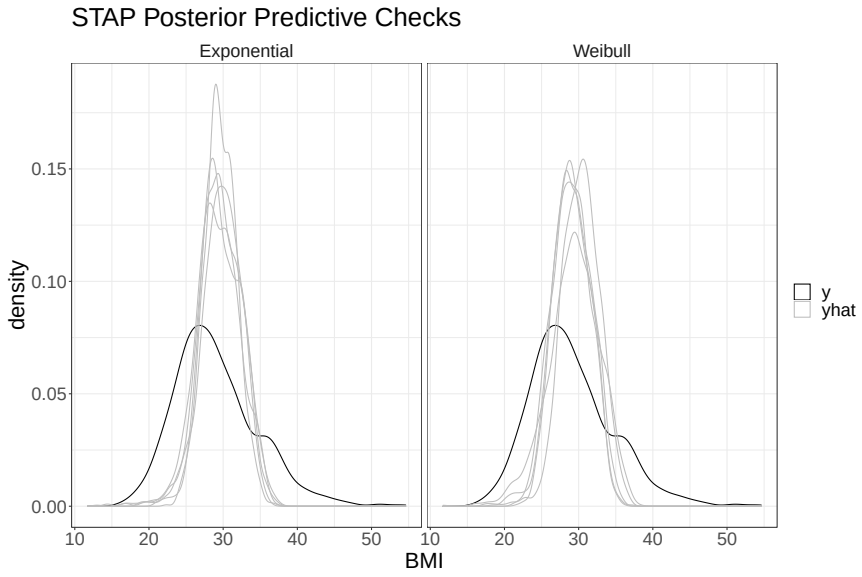


Figure A.5: Posterior Predictive Checks two STAP estimated models. The dark line indicates the observed marginal density estimate, while the gray lines are samples from the estimated posterior predictive distribution.

A.2 Chapter 3 Supplementary Materials

Estimation Details: There is typically a centering constraint enforced on the basis functions to ensure identifiability between the intercept in $\boldsymbol{X}_i$ and the intercept of the basis function expansion. This is not required for our approach as the outer sum in (2) (in the main text) results in a column with the values of $\mid \mathcal{D}_i \mid$ instead of an intercept, and it is generally true in our work with BEF data that $\mid \mathcal{D}_i \mid$ – e.g. the number of FFRs within the inclusion distance for subject i - varies greatly across subjects. This is consistent with the standard treatment for linear functionals of nonlinear functions in regression models (*Wood*, 2017).

Note in all notation that follows, variables with a $\star$ superscript indicate that they been adjusted to remove zero or low member clusters. E.g. $K^\star$, is the number of total non-zero mixture components such that $K^\star \leq K$.

Closed form posterior distributions

$$
\begin{aligned}
p(\boldsymbol{\beta}^{\star}|-) &\sim \text{MVN}_{L+p}(\boldsymbol{V}^{\star}\boldsymbol{X}^{\star,T}\boldsymbol{y}, \boldsymbol{V}^{\star}) \qquad \text{(Conditional Posteriors)} \\
\boldsymbol{V}^{\star} &:= (\boldsymbol{X}^{\star,T}\boldsymbol{X}^{\star} + \boldsymbol{\Lambda}^{\star})^{-1} \\
p(\sigma^{-2}|-) &\propto \text{Gamma}(a_{\sigma^{-2}} + \frac{N}{2} + \frac{L * K^{\star}}{2}, 1 + b_{\sigma^{-2}} + s) \\
s &= \frac{1}{2}\left(|| \, y - \boldsymbol{X}^{\star}\boldsymbol{\beta}^{\star} \, ||^2 + \boldsymbol{\beta}^{\star}\Lambda^{\star}\boldsymbol{\beta}^{\star}\right) \\
p(\alpha|-) &\propto \text{Gamma}(N + a_{\alpha} - 1, b_{\alpha} - \log(1 - \sum_{k=1}^{K-1} u_k)) \\
p(\tau|-) &\propto \text{Gamma}(a_{\tau} + \frac{(L-\mu)}{2}, b_{\tau} + \frac{|| \, \boldsymbol{\beta}^{\star}_{z=1} \, ||^2}{2}) \\
p(v_k|-) &\propto \text{Beta}(1 + n_k, \alpha + \sum_{t=k+1}^{K} n_k)
\end{aligned}
$$

Result: Samples from Posterior

initialization;

while *While sampling* **do**

1. Sample Cluster Labels $\zeta_i^k \propto \mathcal{N}(y - \boldsymbol{X}\boldsymbol{\gamma} - \Phi(d)\boldsymbol{\beta}_k, \sigma^2)$
2. Update Sample Weights draw v_k from (Conditional Posteriors)
3. Create design matrix $\boldsymbol{X}^\star$;
   - Ensure no zero columns as a result of zero-member clusters
4. Solve $\beta^\star = (\boldsymbol{X}^{\star,T}\boldsymbol{X}^\star + \boldsymbol{\Lambda}^\star)^{-1}$
5. draw $\sigma^{-2}, \boldsymbol{\tau}$ from (Conditional Posteriors)

if *zero-member clusters* **then**
  sample corresponding $\boldsymbol{\beta}, \boldsymbol{\tau}$ parameters from prior

end

end

Algorithm 1: Gibbs Sampling Algorithm for drawing samples for STAP-DP model. $\boldsymbol{X}^\star$ is defined as the design matrix consisting of $\boldsymbol{X}$ and Φ combined, where rows of $\boldsymbol{\Phi}$ will be spread out across several columns according to the number of clusters estimated via the DP. $\boldsymbol{\Lambda}$ is the matrix containing penalty parameters $(\tau_{1,k}, \tau_{2,k})$ for each kth DP component.

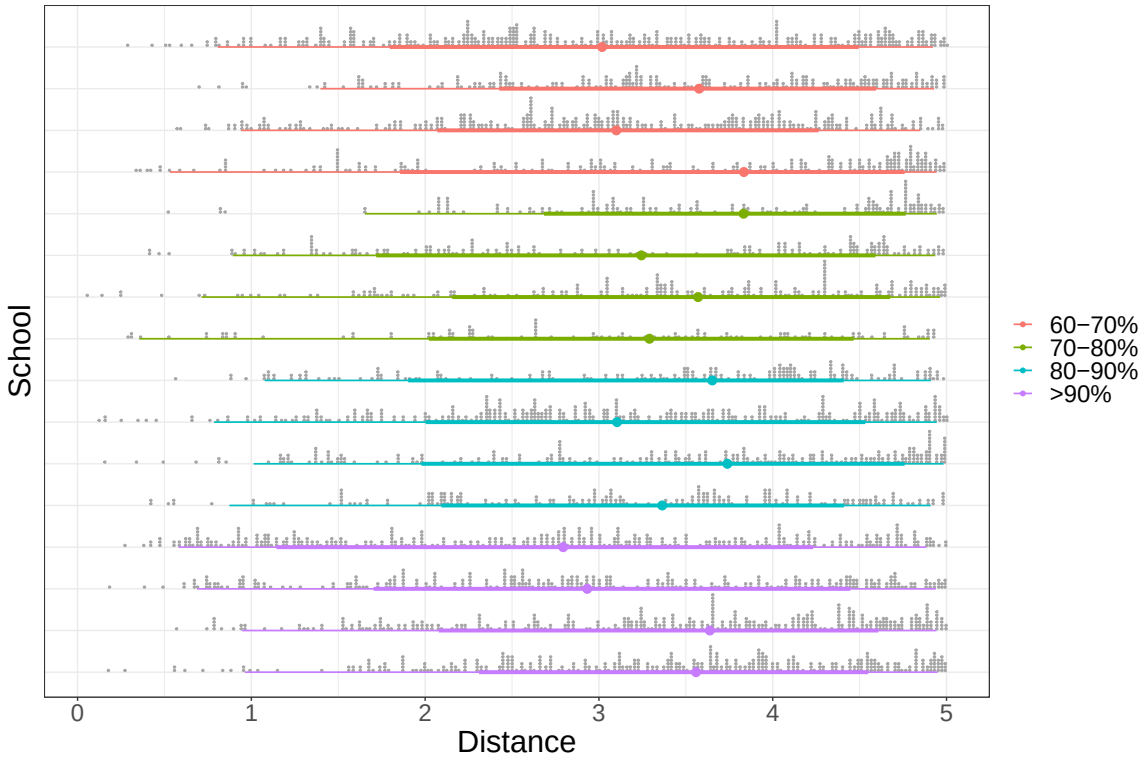


Figure A.6: School level distribution of school-Fast Food distances for calendar year 2001 amongst schools with their highest co-clustering probabilities. The thin line span represents the 2.5 and 97.5 % of the Distance distribution, thick line represents the 50% interval, while the point location represents the median distance. Points above the line represent a quantile dot histogram - see *Fernandes et al.* (2018). Lines are sorted by their highest co-clustering probability category.

Sensitivity Analysis

Prior	Min	1st Quartile	Median	Mean	3rd Quartile	Max
Gamma(1,1)	2	2	2	2	2	2
Gamma(10,10)	5	5	5	5	5	6

Table A.2: Distribution of the number of clusters using two different prior distributions for the concentration parameter in the obesity study among children in Los Angeles. The gamma(1,1) is the prior used for the primary results shown in the chapter.

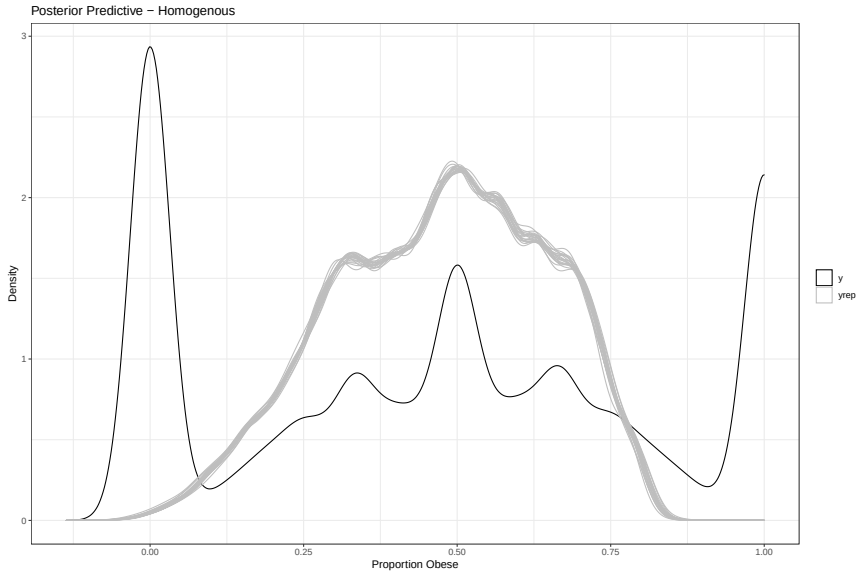


Figure A.7: Posterior Predictive Checks for Homogeneous STAP model. The dark line indicates the observed marginal density estimate, while the gray lines are samples from the estimated posterior predictive distribution.

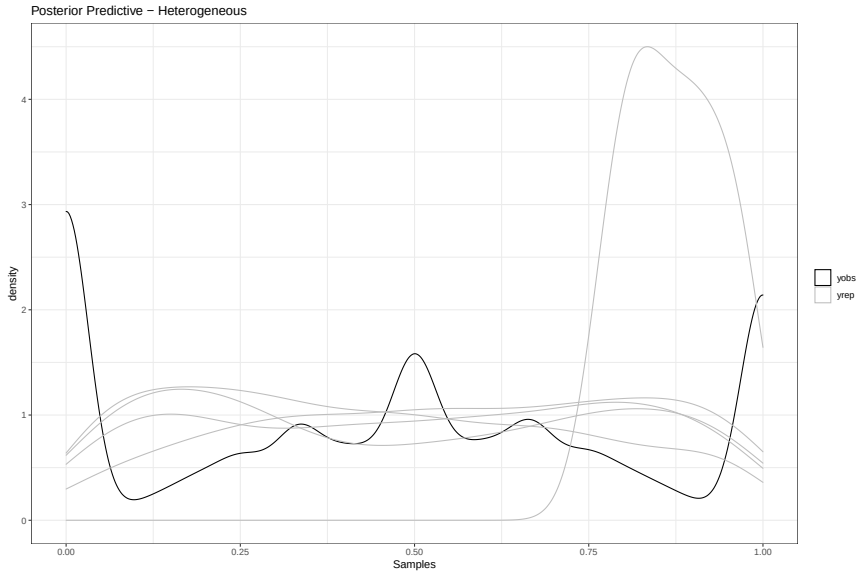


Figure A.8: Posterior Predictive Checks for Heterogeneous STAP model. The dark line indicates the observed marginal density estimate, while the gray lines are samples from the estimated posterior predictive distribution.

A.3 Chapter 4 Supplementary Materials

		Mode Cluster							
		0 (n=426)	1 (n=103)	2 (n=28)	3 (n=231)	4 (n=105)	5 (n=252)	6 (n=31)	Total (n=1176)
FFR Quantity within 1 mile	[1,4]	0	42	39	48	34	47	55	29
	≥5	0	58	61	52	66	53	45	35
	Zero	100	0	0	0	0	0	0	36
Urbanicity	Rural	39	10	14	13	10	6	19	21
	Sub-Urban	34	40	39	44	44	46	42	40
	Urban	27	50	46	43	47	48	39	39
Majority Race/ethnicity	African American	2	2	4	1	0	0	3	1
among enrolled students	Asian	2	6	4	4	4	5	3	4
	Hispanic	27	38	21	27	29	29	16	28
	No Majority	9	11	14	13	18	13	19	12
	White	59	44	57	55	50	52	58	55
Median Income (1,000 USD)	Median	53.9	55.2	55.7	61.0	69.7	61.1	67.4	58.6
	(Q1-Q3)	(41.8-76)	(44-77.2)	(45.4-75.7)	(44.1-83.2)	(48.9-90.6)	(47-77)	(43.4-86.5)	(44.2-79.3)
	IQR	34.1	33.3	30.2	39.1	41.7	30.1	43.0	35.1
Proportion of adults	Median	24.9	25.0	25.4	25.4	25.6	25.3	25.5	25.2
with ≥ 16 years of education	(Q1-Q3)	(24.1-26.2)	(24.2-26.9)	(24.1-27.5)	(24.4-27)	(24.3-27.2)	(24.1-26.7)	(24.5-26.9)	(24.2-26.6)
	IQR	2.1	2.7	3.4	2.6	2.9	2.6	2.5	2.4

Table A.3: Descriptive statistics of school characteristics by mode cluster assignment. Summary statistics - percents, median and inter-quartile range (IQR) for categorical and continuous school-level or census-tract level covariates for each cluster. In the table, the column designated as "Cluster 0" reports summary statistics for those high schools without any fast food restaurants within one mile of their location."Median Income" and "Proportion of residents" refer to characteristics of the population living in the census tract in which schools are located.

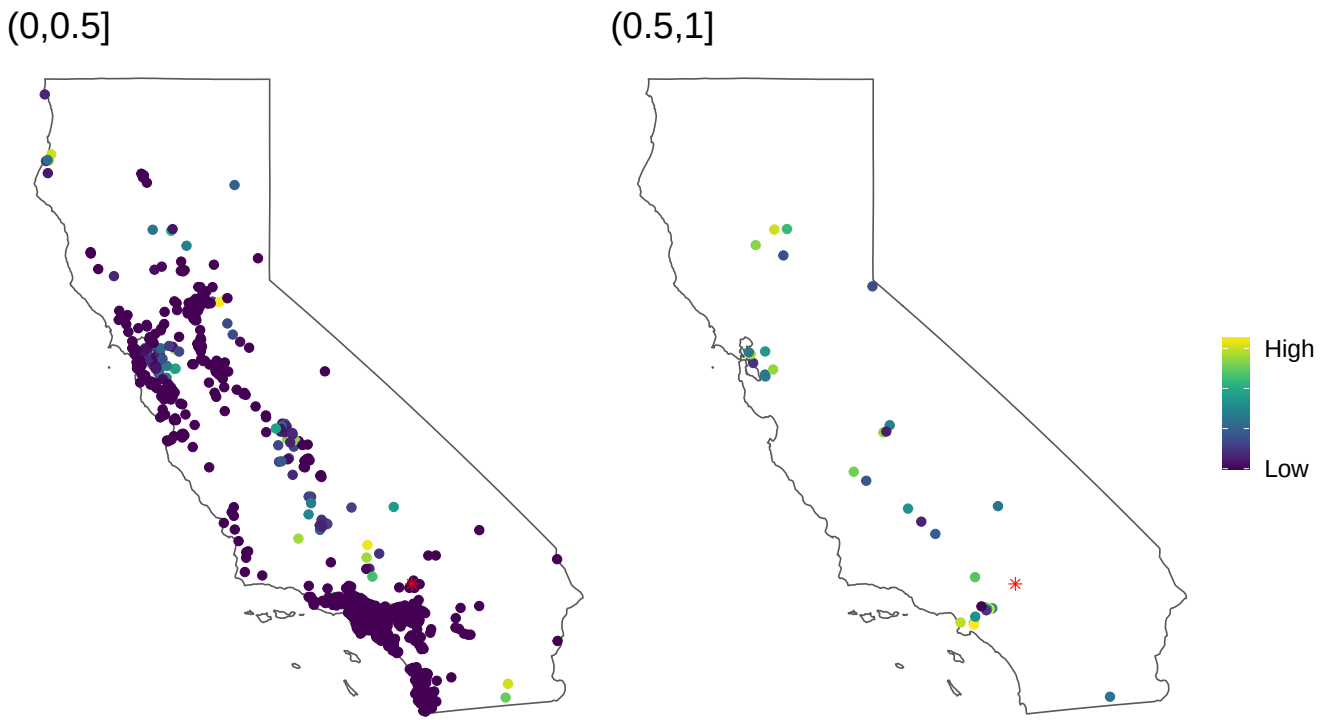


Figure A.9: Map of the probability of co-clustering with the school denoted by a star. Probabilities are color-coded with lighter colors indicating larger probabilities within each of 2 probability intervals considered, (0,0.5] and (0.5,1].

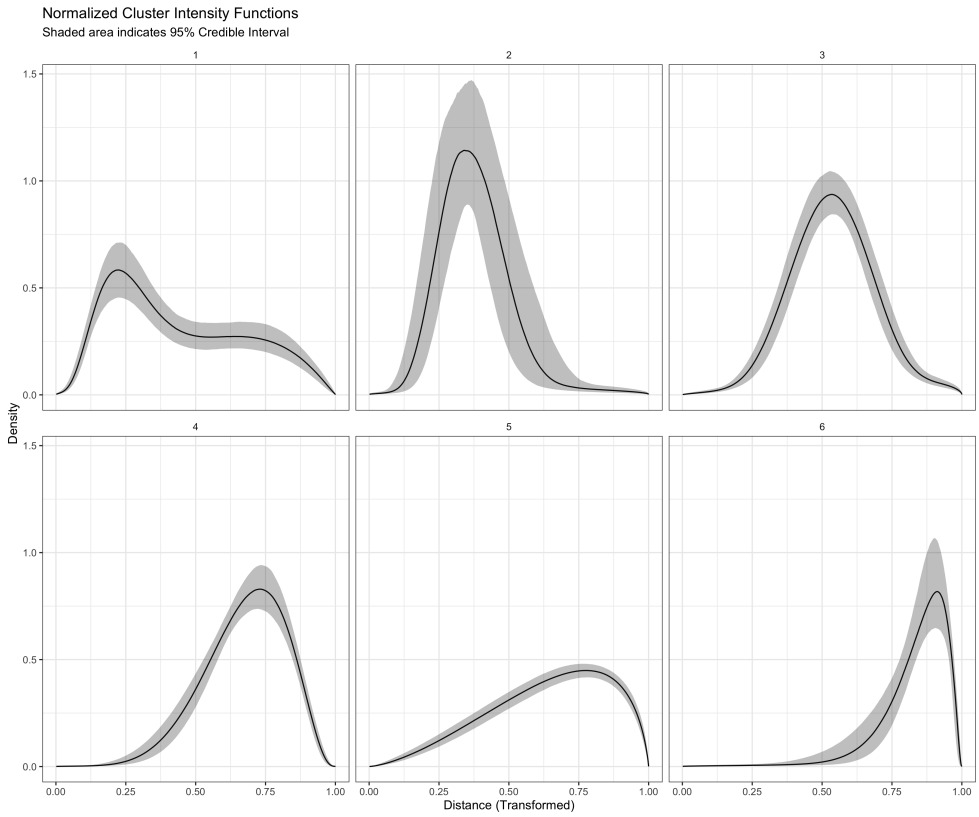


Figure A.10: Median and 95% Credible Interval Estimates for cluster normalized intensity functions on transformed $\mathbb{R}$ scale.

BKMR and Mode GLM Results on the Full Dataset

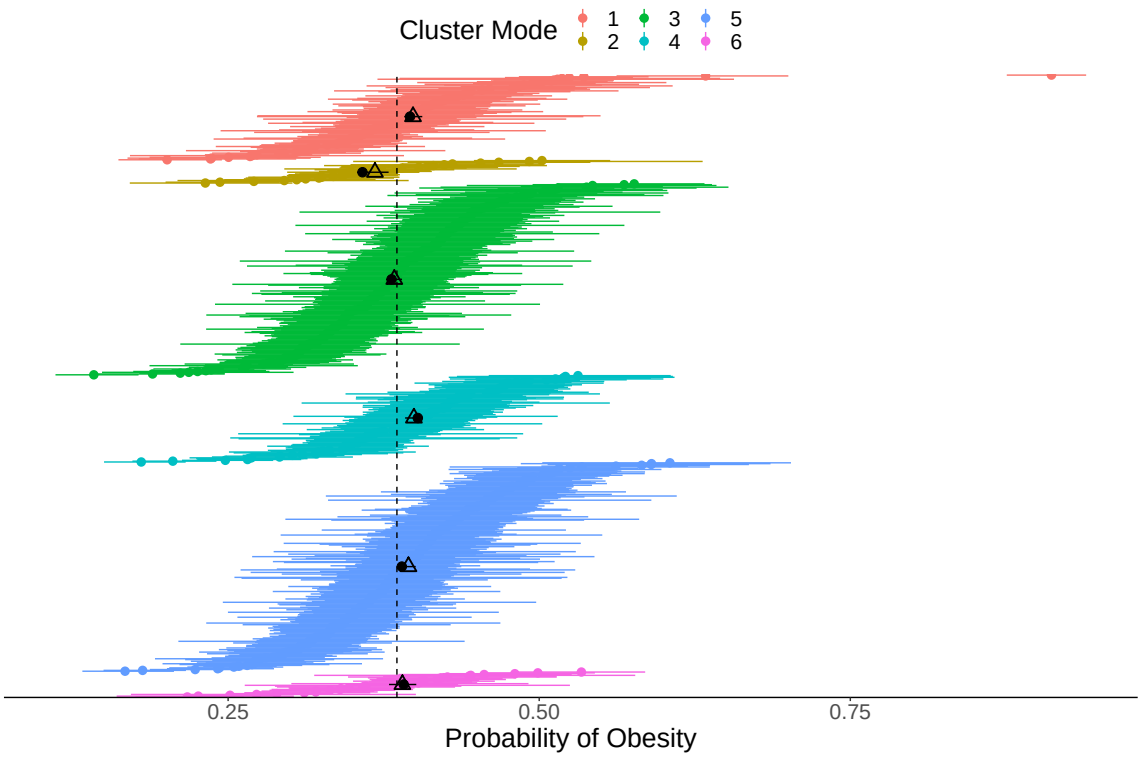


Figure A.11: Health Outcome Fast Food Restaurant (FFR) Spatial Proximity Effects. Bayesian Kernel Machine Regression (BKMR) random school intercepts 95 % credible interval are plotted as lines with colored cluster median dots. Mode GLM (MGLM) effects' 95% credible intervals for each cluster are plotted with triangles denoting the median estimate. The reference dotted line is the posterior mean probability of obesity for children in suburban high schools with a majority of white students, with at least one FFR within a mile of the school's location. BKMR and MGLM results are estimated from a datasets of 1176 schools.

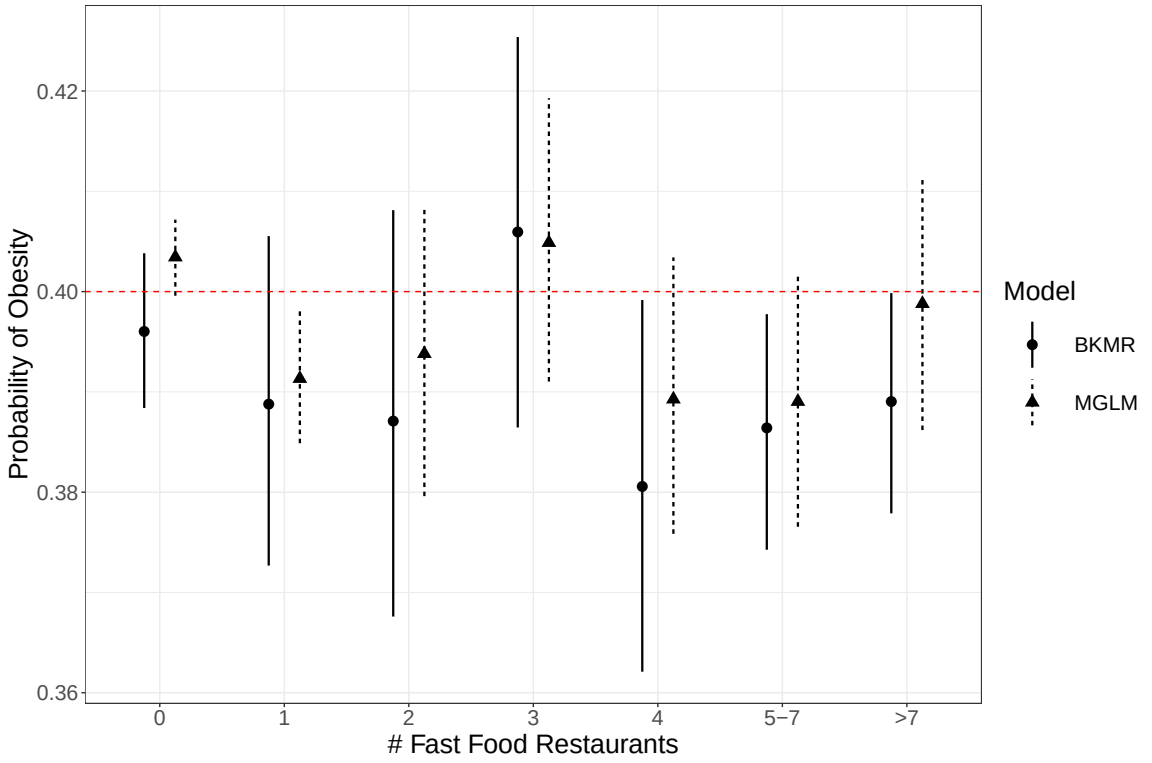


Figure A.12: Health Outcome fast food restaurant Quantity Effect from full dataset. Point ranges represent median and 95% credible intervals

Mode vs. Consensus GLM

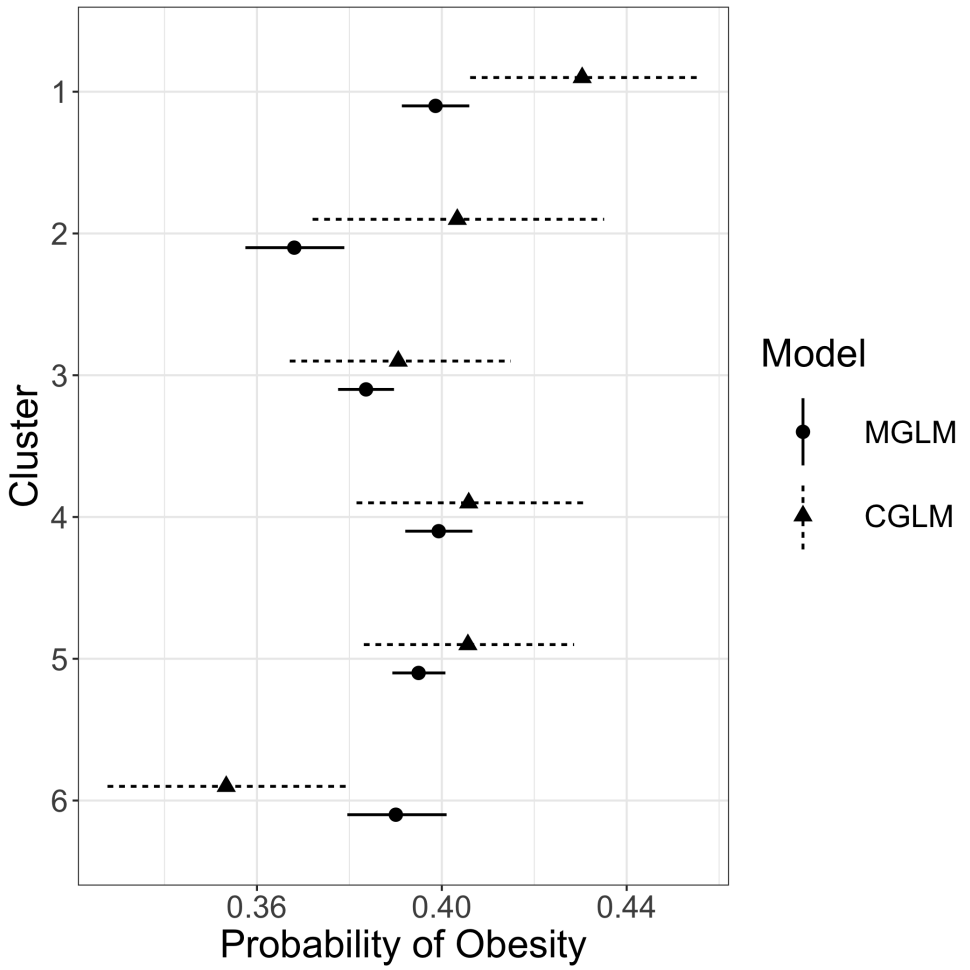


Figure A.13: Posterior Mode (MGLM) and Consensus GLM (CGLM) analyses. Results show the school's proportion of obese students within each cluster configuration.

Models	Full	Consensus
T.1	30,566.95	12,217.12
T.2	33,972.94	17,040.08
T.3	26,096.42	12,169.42
BKMR	11,126.52	6,922.43
CGLM	-	9,883.43
MGLM	17,612.0	-

Table A.4: Widely Applicable Information Criterion (WAIC) *Vehtari et al.* (2017) for Traditional (T) models 1-3, Bayesian Kernel Machine Regression (BKMR) and Consensus GLM (CGLM) for both Consensus and Full datasets corresponding to "In Consensus" and "All" columns from Table 3, respectively. Each model contains the same adjusting covariates and different measures of FFR exposure in a logistic regression modeling 9th grader obesity. T. 1 includes the # of FFR within 1 mile of the school. T. 2 includes the distance to the closest FFR and T. 3 includes both the previous measures. CGLM, MGLM and BKMR are as denoted in the text.

	≥ 1 FF	no FFRs nearby	All schools
% FRPM	48.9 (26.1-68)	48.9 (27.6-71.9)	48.9 (26.9-68.8)
Traditional High School	78.13	78.25	78.18
Charter High School	21.87	21.75	21.82

Table A.5: Supplemental School Covariate Information. The upper half of the table contains the Median (25% quartile,75%quartile) of schools' proportion of students receiving free or reduced price meals. The lower half of the table contains the column percentage of schools that are either Charter or traditional.

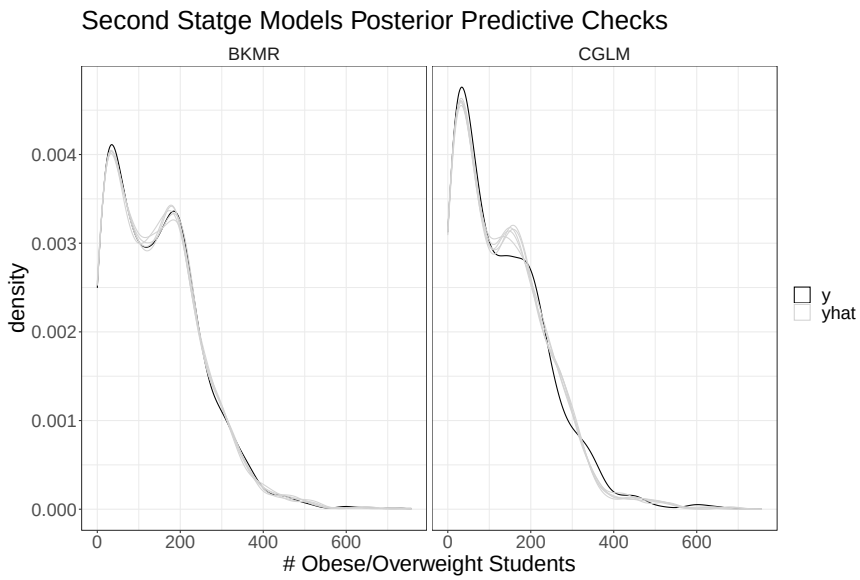


Figure A.14: Posterior Predictive Checks for the two health outcome models. The dark line indicates the observed marginal density estimate, while the gray lines are samples from the estimated posterior predictive distribution.

BIBLIOGRAPHY

BIBLIOGRAPHY

An, R., and R. Sturm (2012), School and residential neighborhood food environment and diet among California youth, *American journal of preventive medicine*, *42*(2), 129–135.

Auchincloss, A. H., A. V. D. Roux, D. G. Brown, C. A. Erdmann, and A. G. Bertoni (2008), Neighborhood resources for physical activity and healthy foods and their association with insulin resistance, *Epidemiology*, pp. 146–157.

Auchincloss, A. H., K. A. Moore, L. V. Moore, and A. V. D. Roux (2012), Improving retrospective characterization of the food environment for a large region in the United States during a historic time period, *Health & place*, *18*(6), 1341–1347.

Austin, S. B., S. J. Melly, B. N. Sanchez, A. Patel, S. Buka, and S. L. Gortmaker (2005), Clustering of fast-food restaurants around schools: a novel application of spatial statistics to the study of food environments, *American journal of public health*, *95*(9), 1575–1581.

Baek, J., B. N. Sánchez, V. J. Berrocal, and E. V. Sanchez-Vaznaugh (2016a), Distributed lag models: examining associations between the built environment and health, *Epidemiology (Cambridge, Mass.)*, *27*(1), 116.

Baek, J., E. V. Sanchez-Vaznaugh, and B. N. Sánchez (2016b), Hierarchical distributed-lag models: exploring varying geographic scale and magnitude in associations between the built environment and health, *American journal of epidemiology*, *183*(6), 583–592.

Baek, J., J. A. Hirsch, K. Moore, L. P. Tabb, T. Barrientos-Gutierrez, L. D. Lisabeth, A. V. Diez-Roux, and B. N. Sánchez (2017), Methods to study variation in the associations between food store availability and body mass in the multi-ethnic study of atherosclerosis, *Epidemiology (Cambridge, Mass.)*, *28*(3), 403.

Bandeen-Roche, K., C. B. Hall, W. F. Stewart, and S. L. Zeger (1999), Modelling disease progression in terms of exposure history, *Statistics in medicine*, *18*(21), 2899–2916.

Bandeen-Roche, K., T. Glass, and K. Bolla (2010), Cumulative lead dose and cognitive function in older adults, *Alternative Medicine Review*, *15*(2), 112–113.

Barrientos-Gutierrez, T., K. A. Moore, A. H. Auchincloss, M. S. Mujahid, C. August, B. N. Sanchez, and A. V. Diez Roux (2017), Neighborhood physical environment and changes in body mass index: results from the Multi-Ethnic Study of Atherosclerosis, *American Journal of Epidemiology*, *186*(11), 1237–1245.

Besser, L. M., D. A. Rodriguez, N. McDonald, W. A. Kukull, A. L. Fitzpatrick, S. R. Rapp, and T. Seeman (2018), Neighborhood built environment and cognition in non-demented older adults: the multi-ethnic study of atherosclerosis, *Social Science & Medicine*, *200*, 27–35.

Bild, D. E., et al. (2002), Multi-ethnic study of atherosclerosis: objectives and design, *American journal of epidemiology*, *156*(9), 871–881.

Binder, D. A. (1978), Bayesian cluster analysis, *Biometrika*, *65*(1), 31–38.

Bobb, J. F., L. Valeri, B. Claus Henn, D. C. Christiani, R. O. Wright, M. Mazumdar, J. J. Godleski, and B. A. Coull (2015), Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures, *Biostatistics*, *16*(3), 493–508.

Bojorquez, I., and L. Ojeda-Revah (2018), Urban public parks and mental health in adult women: Mediating and moderating factors, *International Journal of Social Psychiatry*, *64*(7), 637–646.

Boone-Heinonen, J., P. Gordon-Larsen, C. I. Kiefe, J. M. Shikany, C. E. Lewis, and B. M. Popkin (2011), Fast food restaurants and food stores: longitudinal associations with diet in young to middle-aged adults: the cardia study, *Archives of internal medicine*, *171*(13), 1162–1170.

Booth, K. M., M. M. Pinkston, and W. S. C. Poston (2005), Obesity and the built environment, *Journal of the American Dietetic Association*, *105*(5), 110–117.

Camerlenghi, F., D. B. Dunson, A. Lijoi, I. Prünster, A. Rodríguez, et al. (2019), Latent nested nonparametric priors (with discussion), *Bayesian Analysis*, *14*(4), 1303–1356.

Carpenter, B., et al. (2016), Stan: A probabilistic programming language, *Journal of Statistical Software*, *20*(2), 1–37.

Chaix, B., J. Merlo, S. Subramanian, J. Lynch, and P. Chauvin (2005), Comparison of a spatial perspective with the multilevel analytical approach in neighborhood studies: the case of mental and behavioral disorders due to psychoactive substance use in Malmö, Sweden, 2001, *American journal of epidemiology*, *162*(2), 171–182.

Charreire, H., R. Casey, P. Salze, C. Simon, B. Chaix, A. Banos, D. Badariotti, C. Weber, and J.-M. Oppert (2010), Measuring the food environment using geographical information systems: a methodological review, *Public health nutrition*, *13*(11), 1773–1785.

Chiang, S., M. Guindani, H. J. Yeh, S. Dewar, Z. Haneef, J. M. Stern, and M. Vannucci (2017), A hierarchical Bayesian model for the identification of PET markers associated to the prediction of surgical outcome after anterior temporal lobe resection, *Frontiers in neuroscience*, *11*, 669.

Coker, E., J. Chevrier, S. Rauch, A. Bradman, M. Obida, M. Crause, R. Bornman, and B. Eskenazi (2018), Association between prenatal exposure to multiple insecticides and child body weight and body composition in the VHEMBE South African birth cohort, *Environment international*, *113*, 122–132.

Cook, S. R., A. Gelman, and D. B. Rubin (2006), Validation of software for Bayesian models using posterior quantiles, *Journal of Computational and Graphical Statistics*, *15*(3), 675–692.

Currie, J., S. DellaVigna, E. Moretti, and V. Pathania (2010), The effect of fast food restaurants on obesity and weight gain, *American Economic Journal: Economic Policy*, *2*(3), 32–63.

Davis, B., and C. Carpenter (2009), Proximity of fast-food restaurants to schools and adolescent obesity, *American Journal of Public Health*, *99*(3), 505–510.

Diebolt, J., and C. P. Robert (1994), Estimation of finite mixture distributions through bayesian sampling, *Journal of the Royal Statistical Society: Series B (Methodological)*, *56*(2), 363–375.

Ding, D., and K. Gebel (2012), Built environment, physical activity, and obesity: what have we learned from reviewing the literature?, *Health & place*, *18*(1), 100–105.

Eilers, P. H., and B. D. Marx (1996), Flexible smoothing with b-splines and penalties, *Statistical science*, pp. 89–102.

Evenson, K. R., S. A. Jones, K. M. Holliday, D. A. Cohen, and T. L. McKenzie (2016), Park characteristics, use, and physical activity: A review of studies using SOPARC (system for observing play and recreation in communities), *Preventive medicine*, *86*, 153–166.

Ferguson, T. S. (1973), A bayesian analysis of some nonparametric problems, *The annals of statistics*, pp. 209–230.

Fernandes, M., L. Walls, S. Munson, J. Hullman, and M. Kay (2018), Uncertainty displays using quantile dotplots or cdfs improve transit decision-making, in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–12.

Ferrari, D. (2020), Modeling context-dependent latent effect heterogeneity, *Political Analysis*, *28*(1), 20–46.

Fitzmaurice, G., M. Davidian, G. Verbeke, and G. Molenberghs (2008), *Longitudinal data analysis*, CRC press.

Fotheringham, A. S., and D. W. Wong (1991), The modifiable areal unit problem in multivariate statistical analysis, *Environment and planning A*, *23*(7), 1025–1044.

Friedman, J., T. Hastie, and R. Tibshirani (2001), *The elements of statistical learning*, vol. 1, Springer series in statistics New York.

Garin, N., B. Olaya, M. Miret, J. L. Ayuso-Mateos, M. Power, P. Bucciarelli, and J. M. Haro (2014), Built environment and elderly population health: a comprehensive literature review, *Clinical practice and epidemiology in mental health: CP & EMH*, *10*, 103.

Gelman, A., and J. Hill (2007), Data analysis using regression and hierarchical/multilevel models, *New York, NY: Cambridge*.

Gelman, A., H. S. Stern, J. B. Carlin, D. B. Dunson, A. Vehtari, and D. B. Rubin (2013), *Bayesian data analysis*, Chapman and Hall/CRC.

Gelman, A., D. Simpson, and M. Betancourt (2017), The prior can often only be understood in the context of the likelihood, *Entropy*, *19*(10), 555.

Graziani, R., M. Guindani, and P. F. Thall (2015), Bayesian nonparametric estimation of targeted agent effects on biomarker change to predict clinical outcome, *Biometrics*, *71*(1), 188–197.

Guo, J. Y., and C. R. Bhat (2004), Modifiable areal units: problem or perception in modeling of residential location choice?, *Transportation Research Record*, *1898*(1), 138–147.

Guo, Y., A. G. Barnett, X. Pan, W. Yu, and S. Tong (2011), The impact of temperature on mortality in Tianjin, China: a case-crossover design with a distributed lag nonlinear model, *Environmental health perspectives*, *119*(12), 1719.

Hartigan, J. A. (1975), *Clustering algorithms*, John Wiley & Sons, Inc.

Heaton, M. J., and A. E. Gelfand (2011), Spatial regression using kernel averaged predictors, *Journal of agricultural, biological, and environmental statistics*, *16*(2), 233–252.

Heaton, M. J., and A. E. Gelfand (2012), Kernel averaged predictors for spatio-temporal regression models, *Spatial statistics*, *2*, 15–32.

Hirsch, J. A., A. V. Diez-Roux, K. A. Moore, K. R. Evenson, and D. A. Rodriguez (2014), Change in walking and body mass index following residential relocation: the multi-ethnic study of atherosclerosis, *American journal of public health*, *104*(3), e49–e56.

Hoehner, C. M., and M. Schootman (2010), Concordance of commercial data sources for neighborhood-effects studies, *Journal of Urban Health*, *87*(4), 713–725.

Hoffman, M. D., and A. Gelman (2014), The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo., *Journal of Machine Learning Research*, *15*(1), 1593–1623.

Howard, P. H., M. Fitzpatrick, and B. Fulfrost (2011), Proximity of food retailers to schools and rates of overweight ninth grade students: an ecological study in California, *BMC Public Health*, *11*(1), 68.

Institute, T. C. (2001), The fitnessgram assessment, accessed 2020.

Ishwaran, H., and L. F. James (2001), Gibbs sampling methods for stick-breaking priors, *Journal of the American Statistical Association*, *96*(453), 161–173.

James, P., D. Berrigan, J. E. Hart, J. A. Hipp, C. M. Hoehner, J. Kerr, J. M. Major, M. Oka, and F. Laden (2014), Effects of buffer size and shape on associations between the built environment and energy balance, *Health & place*, *27*, 162–170.

Ji, C., D. Merl, T. B. Kepler, and M. West (2009), Spatial mixture modelling for unobserved point processes: Examples in immunofluorescence histology, *Bayesian analysis (Online)*, *4*(2), 297.

Kaiser, P., A. V. Diez Roux, M. Mujahid, M. Carnethon, A. Bertoni, S. D. Adar, S. Shea, R. McClelland, and L. Lisabeth (2016), Neighborhood environments and incident hypertension in the multi-ethnic study of atherosclerosis, *American journal of epidemiology*, *183*(11), 988–997.

Kaufman, T. K., D. M. Sheehan, A. Rundle, K. M. Neckerman, M. D. Bader, D. Jack, and G. S. Lovasi (2015), Measuring health-relevant businesses over 21 years: refining the National Establishment Time-Series (NETS), a dynamic longitudinal data set, *BMC research notes*, *8*(1), 507.

Kaufman, T. K., A. Rundle, K. M. Neckerman, D. M. Sheehan, G. S. Lovasi, and J. A. Hirsch (2019), Neighborhood recreation facilities and facility membership are jointly associated with objectively measured physical activity, *Journal of urban health*, *96*(4), 570–582.

Kern, D. M., A. H. Auchincloss, M. F. Stehr, A. V. Diez Roux, L. V. Moore, G. P. Kanter, and L. F. Robinson (2017), Neighborhood prices of healthier and unhealthier foods and associations with diet quality: Evidence from the multi-ethnic study of atherosclerosis, *International journal of environmental research and public health*, *14*(11), 1394.

Kwan, M.-P. (2013), Beyond space (as we knew it): Toward temporally integrated geographies of segregation, health, and accessibility: Space–time integration in geography and giscience, *Annals of the Association of American Geographers*, *103*(5), 1078–1086.

Kwan, M.-P. (2018), The limits of the neighborhood effect: Contextual uncertainties in geographic, environmental health, and social science research, *Annals of the American Association of Geographers*, *108*(6), 1482–1490.

Lau, J. W., and P. J. Green (2007), Bayesian model-based clustering procedures, *Journal of Computational and Graphical Statistics*, *16*(3), 526–558.

Leal, C., K. Bean, F. Thomas, and B. Chaix (2011), Are associations between neighborhood socioeconomic characteristics and body mass index or waist circumference based on model extrapolations?, *Epidemiology*, pp. 694–703.

Liu, R., R. Giordano, M. I. Jordan, and T. Broderick (2018), Evaluating sensitivity to the stick breaking prior in bayesian nonparametrics, *arXiv preprint arXiv:1810.06587*.

MacEachern, S. N. (2000), Dependent dirichlet processes, *Unpublished manuscript, Department of Statistics, The Ohio State University*, pp. 1–40.

MacEachern, S. N., and X. Shen (1999), Variable selection and function estimation in additive nonparametric regression using a data-based prior: Comment, *Journal of the American Statistical Association*, *94*(447), 799–802.

Macintyre, S., A. Ellaway, and S. Cummins (2002), Place effects on health: how can we conceptualise, operationalise and measure them?, *Social Science & Medicine*, *55*(1), 125–139, doi:https://doi.org/10.1016/S0277-9536(01)00214-3, selected papers from the 9th International Symposium on Medical G eography.

McGuire, S. (2012), Institute of Medicine (IOM) Early Childhood Obesity Prevention Policies. Washington, DC: The National Academies Press; 2011, *Advances in Nutrition*, *3*(1), 56–57, doi:10.3945/an.111.001347.

Miller, J. W. (2014), Nonparametric and variable-dimension bayesian mixture models: Analysis, comparison, and new methods, Ph.D. thesis, Citeseer.

Miller, J. W., and M. T. Harrison (2013), A simple example of dirichlet process mixture inconsistency for the number of components, in *Advances in neural information processing systems*, pp. 199–206.

Miller, J. W., and M. T. Harrison (2014), Inconsistency of Pitman-Yor process mixtures for the number of components, *The Journal of Machine Learning Research*, *15*(1), 3333–3370.

Miller, J. W., and M. T. Harrison (2018), Mixture models with a prior on the number of components, *Journal of the American Statistical Association*, *113*(521), 340–356.

Morgan, S. L. (2013), *Handbook of causal analysis for social research*, Springer.

Neuhaus, J. M., and J. D. Kalbfleisch (1998), Between-and within-cluster covariate effects in the analysis of clustered data, *Biometrics*, pp. 638–645.

Nylund-Gibson, K., R. P. Grimm, and K. E. Masyn (2019), Prediction from latent classes: A demonstration of different approaches to include distal outcomes in mixture models, *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(6), 967–985.

of Education-FitnessGram, C. D. (2017).

Openshaw, S. (1996), Developing GIS-relevant zone-based spatial analysis methods, *Spatial analysis: modelling in a GIS environment*, pp. 55–73.

O'Sullivan, F. (1986), A statistical perspective on ill-posed inverse problems, *Statistical science*, pp. 502–518.

Padgham, M., R. Lovelace, M. Salmon, and B. Rudis (2017), osmdata, *Journal of Open Source Software*, *2*(14).

Papas, M. A., A. J. Alberg, R. Ewing, K. J. Helzlsouer, T. L. Gary, and A. C. Klassen (2007), The built environment and obesity, *Epidemiologic reviews*, *29*(1), 129–143.

Papastamoulis, P. (2016), label.switching: An R package for dealing with the label switching problem in mcmc outputs, *Journal of Statistical Software, Code Snippets*, *69*(1), 1–24, doi:10.18637/jss.v069.c01.

Pedersen, T. L. (2018), Tidygraph: a tidy api for graph manipulation, *R package version*, *1*(0).

Peterson, A. (2020a), bendr: Built environment Nested Dirichlet Processes in R., r package version 0.1.0-alpha.

Peterson, A. (2020b), *rbenvo: Built Environment Objects in R*, version 0.1.0.6.

Peterson, A. (2020c), *rsstap: Spline Spatial Temporal Aggregated Predictors in R*, r package version 0.1.0.

Peterson, A. (2020d), *rstapDP: Functional Dirichlet Process Spatial Temporal Aggregated Predictors in R*, version 0.1.1.

Peterson, A., and B. Sanchez (2018), rstap: An r package for spatial temporal aggregated predictor models, *arXiv preprint arXiv:1812.10208*.

R Core Team (2013), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Raftery, A. E., and S. M. Lewis (1995), The number of iterations, convergence diagnostics and generic metropolis algorithms, *Practical Markov Chain Monte Carlo*, *7*(98), 763–773.

Ray, S., and B. Mallick (2006), Functional clustering by bayesian wavelet methods, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*(2), 305–332.

Ren, L., L. Du, L. Carin, and D. Dunson (2011), Logistic stick-breaking process, *Journal of Machine Learning Research*, *12*(Jan), 203–239.

Renalds, A., T. H. Smith, and P. J. Hale (2010), A systematic review of built environment and health, *Family & community health*, *33*(1), 68–78.

Rodriguez, A., D. B. Dunson, and A. E. Gelfand (2008), The nested dirichlet process, *Journal of the American Statistical Association*, *103*(483), 1131–1154.

Rodriguez, A., D. B. Dunson, et al. (2014), Functional clustering in nested designs: modeling variability in reproductive epidemiology studies, *Annals of Applied Statistics*, *8*(3), 1416–1442.

Rodríguez, C. E., and S. G. Walker (2014), Label switching in bayesian mixture models: Deterministic relabeling strategies, *Journal of Computational and Graphical Statistics*, *23*(1), 25–45.

Roof, K., and N. Oleru (2008), Public health: Seattle and King County's push for the built environment, *Journal of environmental health*, *71*(1), 24–27.

Rose, G. (2001), Sick individuals and sick populations, *International journal of epidemiology*, *30*(3), 427–432.

Roux, A. V. D. (2003), Residential environments and cardiovascular risk, *Journal of Urban Health*, *80*(4), 569–589.

Roux, A. V. D., K. R. Evenson, A. P. McGinn, D. G. Brown, L. Moore, S. Brines, and D. R. Jacobs Jr (2007), Availability of recreational resources and physical activity in adults, *American journal of public health*, *97*(3), 493–499.

Roux, A. V. D., M. S. Mujahid, J. A. Hirsch, K. Moore, and L. V. Moore (2016), The impact of neighborhoods on CV risk, *Global heart*, *11*(3), 353–363.

Sacks, G., B. Swinburn, and G. Xuereb (2012), Population-based approaches to childhood obesity prevention.

San Martın, E., and J. González (2010), Bayesian identifiability: Contributions to an inconclusive debate, *Chilean Journal of Statistics*, *1*(2), 69–91.

Sánchez, B. N., E. V. Sanchez-Vaznaugh, A. Uscilka, J. Baek, and L. Zhang (2012), Differential associations between the food environment near schools and childhood overweight across race/ethnicity, gender, and grade, *American journal of epidemiology*, *175*(12), 1284–1293.

Schipperijn, J., M. Ried-Larsen, M. S. Nielsen, A. F. Holdt, A. Grøntved, A. K. Ersbøll, and P. L. Kristensen (2015), A longitudinal study of objectively measured built environment as determinant of physical activity in young adults: the European Youth Heart Study, *Journal of Physical Activity and Health*, *12*(7), 909–914.

Schoenborn, C. A. (2002), *Body weight status of adults: United States, 1997-98*, 330, US Department of Health and Human Services, Centers for Disease Control and ....

Spielman, S. E., and E.-h. Yoo (2009), The spatial dimensions of neighborhood effects, *Social science & medicine*, *68*(6), 1098–1105.

Stan Development Team (2016), rstanarm: Bayesian applied regression modeling via Stan., r package version 2.13.1.

Stan Development Team (2020), RStan: the R interface to Stan, r package version 2.19.3.

Stephens, M. (2000), Dealing with label switching in mixture models, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *62*(4), 795–809.

Team, S. D. (2017), *Stan Modeling Language Users Guide and Reference Manual*, Version2.17.0.

Valeri, L., et al. (2017), The joint effect of prenatal exposure to metal mixtures on neurodevelopmental outcomes at 20–40 months of age: evidence from rural Bangladesh, *Environmental health perspectives*, *125*(6), 067,015.

Vehtari, A., A. Gelman, and J. Gabry (2017), Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC, *Statistics and Computing*, *27*(5), 1413–1432.

Vehtari, A., A. Gelman, D. Simpson, B. Carpenter, and P.-C. Bürkner (2021), Rank-normalization, folding, and localization: An improved rhat for assessing convergence of mcmc, *Bayesian Analysis*, *1*(1), 1–28.

Wade, S. (2015), *mcclust.ext: Point estimation and credible balls for Bayesian cluster analysis*, r package version 1.0.

Wade, S., Z. Ghahramani, et al. (2018), Bayesian cluster analysis: Point estimation and credible balls (with discussion), *Bayesian Analysis*, *13*(2), 559–626.

Wahba, G. (1990), *Spline models for observational data*, SIAM.

Walker, K., K. Eberwein, and M. Herman (2018), Tidycensus: Load us census boundary and attribute data as' tidyverse'and'sf'-ready data frames, *R package version 0.9*, *6*.

Wall, M. M., and X. Liu (2009), Spatial latent class analysis model for spatially distributed multivariate binary data, *Computational statistics & data analysis*, *53*(8), 3057–3069.

Wall, M. M., N. I. Larson, A. Forsyth, D. C. Van Riper, D. J. Graham, M. T. Story, and D. Neumark-Sztainer (2012), Patterns of obesogenic neighborhood features and adolescent weight: a comparison of statistical approaches, *American journal of preventive medicine*, *42*(5), e65–e75.

Walls, D. (2013), National establishment time-series (nets) database: 2012 database description, *Oakland: Walls & Associates*.

Wang, X., B. Mukherjee, and S. K. Park (2018), Associations of cumulative exposure to heavy metal mixtures with obesity and its comorbidities among US adults in NHANES 2003–2014, *Environment international*, *121*, 683–694.

Wickham, H., and G. Grolemund (2016), *R for data science: import, tidy, transform, visualize, and model data*, " O'Reilly Media, Inc.".

Wickham, H., et al. (2019), Welcome to the tidyverse, *Journal of Open Source Software*, *4*(43), 1686.

Wong, D. (2009), The modifiable areal unit problem (MAUP), *The SAGE handbook of spatial analysis*, *105*(23), 2.

Wood, S. (2017), *Generalized Additive Models: An Introduction with R*, 2 ed., Chapman and Hall/CRC.

Wood, S. N. (2004), Stable and efficient multiple smoothing parameter estimation for generalized additive models, *Journal of the American Statistical Association*, *99*(467), 673–686.

Wood, S. N. (2016), Just another gibbs additive modeller: interfacing jags and mgcv, *arXiv preprint arXiv:1602.02539*.

Xiao, S., A. Kottas, and B. Sansó (2015), Modeling for seasonal marked point processes: An analysis of evolving hurricane occurrences, *The Annals of Applied Statistics*, pp. 353–382.

Zhang, H. (2004), Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics, *Journal of the American Statistical Association*, *99*(465), 250–261.